Программирование и статистический анализ данных на языке R

Лекция 3 (Основы статистического анализа на языке R)

Петровский Михаил (ВМК МГУ), michael@cs.msu.su

Вспомним основные понятия

- Генеральная совокупность => выборка
- Описательные статистики vs статистический вывод
- Доверительные интервалы vs точечные оценки
- Статистический вывод – для проверки гипотез:

ЕСЛИ предположение P истинно, то ВЕРОЯТНО, что Q ложно

по факту наблюдаем, что Q выполняется,

ЗНАЧИТ ВЕРОЯТНО P ложно

А если Q – ложно, что можно сказать про P?

Можно ли использовать разные Q для проверки одной P?

Проверка статистических гипотез

H_o
H_1

Формулируем нулевую и альтернативную гипотезы

Устанавливаем α

Выбираем уровень значимости

Статистическая гипотеза — предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки

Вычисляем статистики и p

$p \geq \alpha \Rightarrow$ Принимаем H_o
$p < \alpha \Rightarrow$ Отвергаем H_o

Принимаем решение

Р-значение равно вероятности того, что случайная величина с данным распределением тестовой статистики при нулевой гипотезе примет значение, более экстремальное, чем фактически полученное значение тестовой статистики (оно же вероятность ложно положительной ошибки, или ошибки I рода)

Уровень значимости и мощность

Реальность / Решение	H_0 Истина	H_0 Ложна
Принимаем H_0	Правильно	Ошибка II рода $p(\text{Type II} \mid H_1)=\beta$
Отвергаем H_0	Ошибка I рода $p(\text{Type I} \mid H_0)=\alpha$	Правильно $(1 - \beta)=$ *Мощность*

Мощность зависит (обратно) от α, размера выборки и зачастую от самой статистики.

Для простых случаев можно напрямую найти необходимый размер выборки при заданных ограничениях на мощность, уровень значимости и в зависимости от проверяемой гипотезы (pwr.<TEST>)

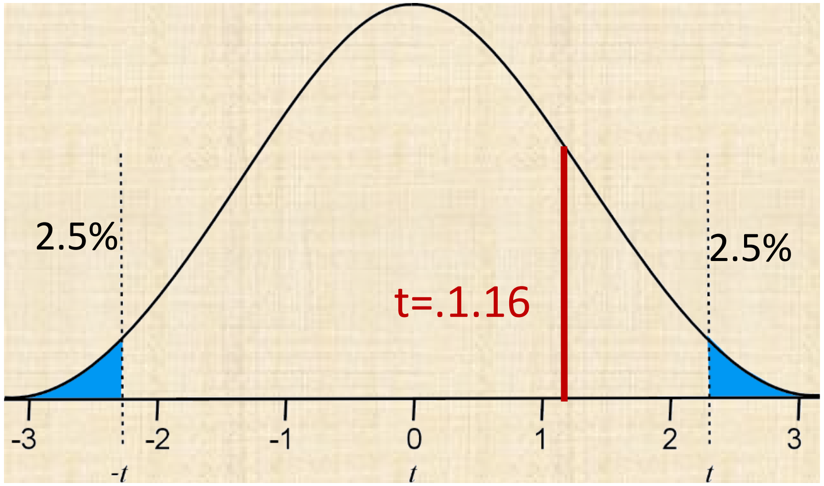
Односторонние и двусторонние тесты.

Процедура TTEST

- Проверяется гипотеза о значении среднего H_0: $\mu=\mu_0$ против H_1: $\mu\neq\mu_0$
- Вычисляется статистика:

$$t = \frac{(\overline{x} - \mu_0)}{s_{\overline{x}}} \qquad S_{\overline{x}} = \frac{S}{\sqrt{n}}$$

- *S* – смещенная оценка дисперсии
- Нулевая гипотеза отвергается если полученные значения «экстримальнее» (как в + так и в -) чем ожидается при заданном уровне значимости



Условия применимости: нормальность, независимость, равенство дисперсий (для многих выборок)

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, …)
```

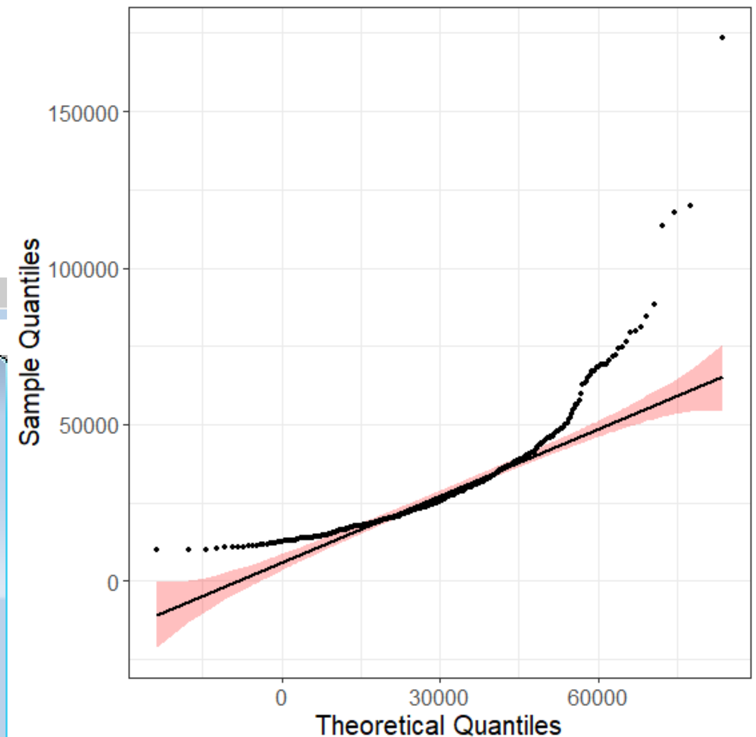
Одновыборочный t-тест

R Console

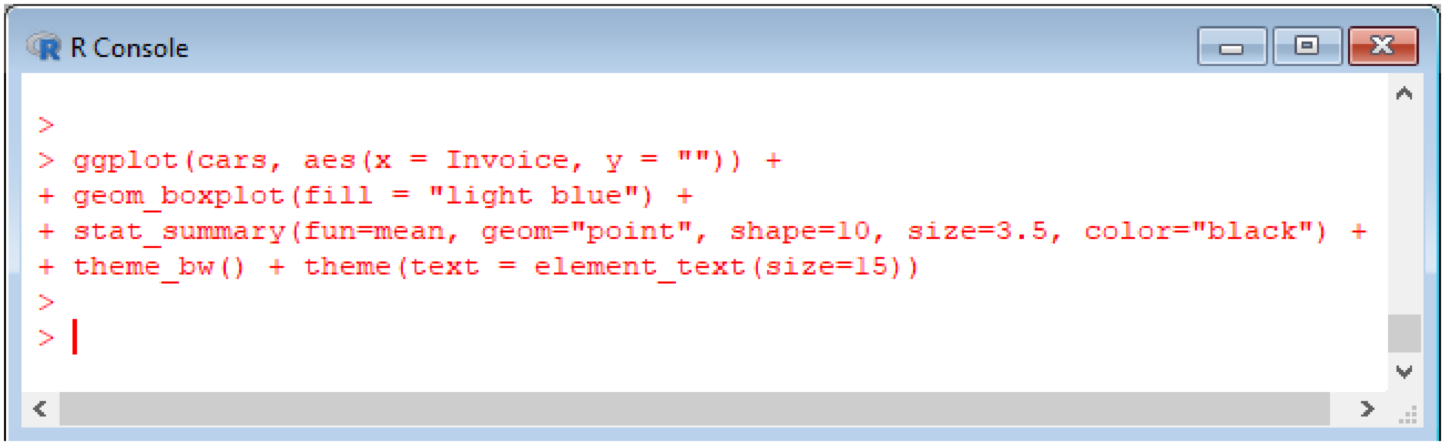
```
> describe(cars$Invoice)
   vars   n    mean       sd  median  trimmed      mad  min    max  range skew kurtosis     se
X1    1 428 30014.7 17642.12 25294.5 27187.34 11165.46 9875 173560 163685 2.81    13.69 852.76
> t.test(cars$Invoice, mu = 30000)

        One Sample t-test

data:  cars$Invoice
t = 0.017239, df = 427, p-value = 0.9863
alternative hypothesis: true mean is not equal to 30000
95 percent confidence interval:
 28338.56 31690.84
sample estimates:
mean of x
  30014.7
```

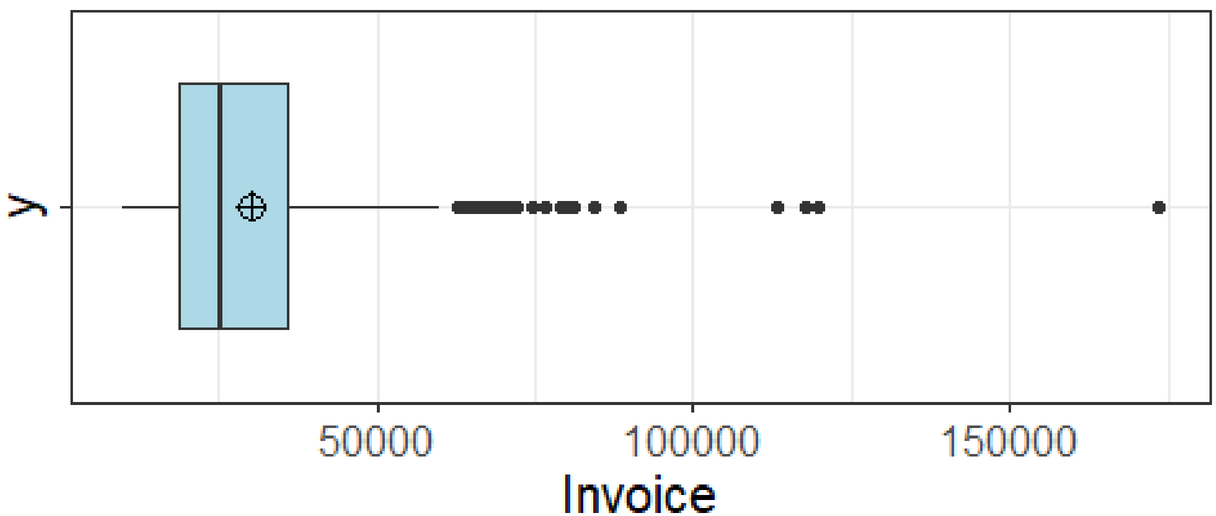
R Console

```
> ggplot(data = cars, mapping = aes(sample = Invoice)) +
+ stat_qq_band(alpha=0.25, conf=0.95, fill="red") +
+ stat_qq_line() +
+ stat_qq_point(col="black", size = 1) +
+ labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
+ theme_bw() + theme(text = element_text(size=15))
> 
```



Одновыборочный t-тест

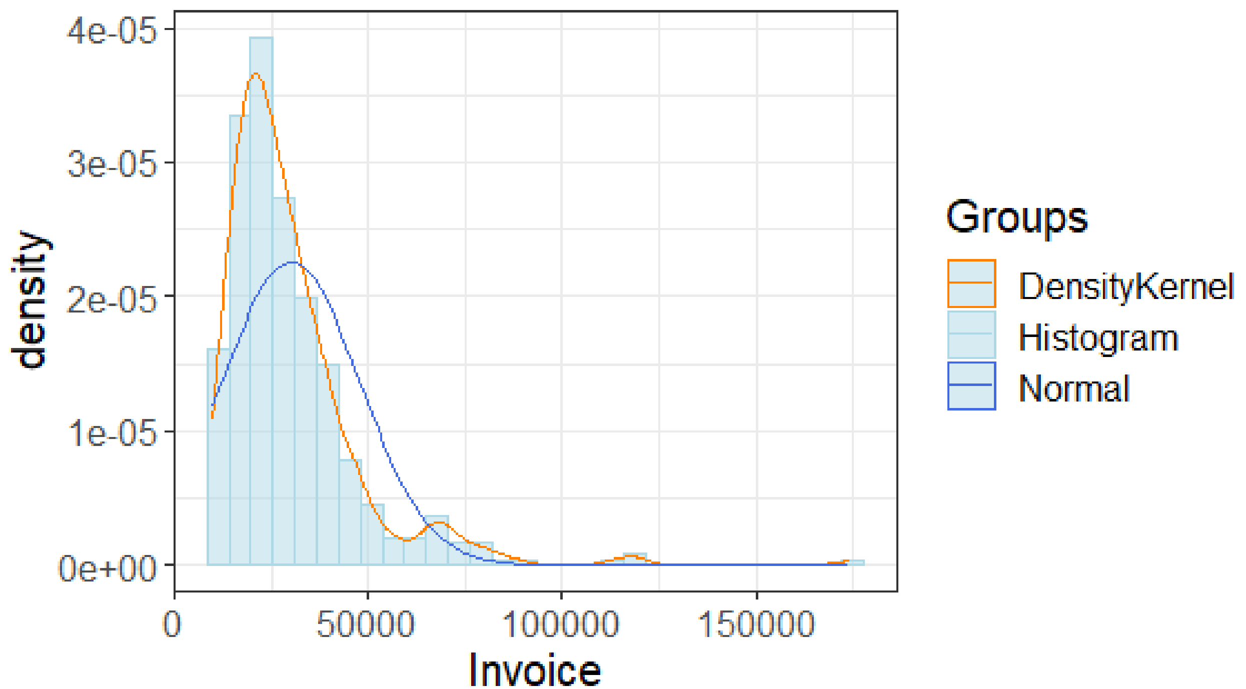




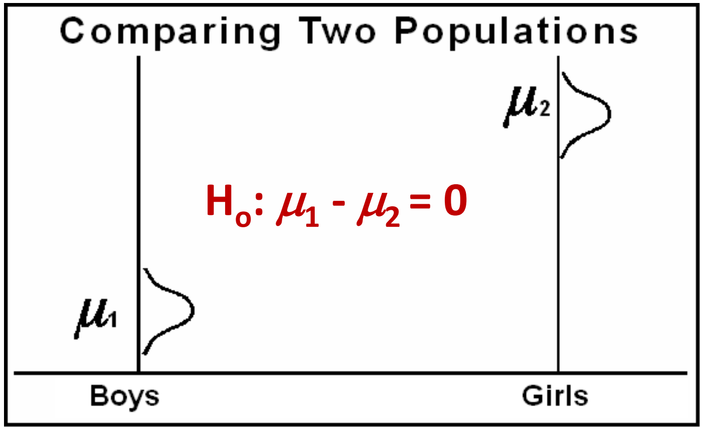
Одновыборочный t-тест

R Console

```
> dnorm_pars <- with(cars, c(mean = mean(Invoice), sd = sd(Invoice)))
> p <- ggplot(cars, aes(x = Invoice)) +
+      geom_histogram(aes(y = ..density.., colour = "Histogram"), fill="light blue", bins=30, alpha=.5) $
+      geom_density(aes(y = ..density.., colour = "DensityKernel"), alpha=.5) +
+      stat_function(fun = dnorm, args = dnorm_pars, aes(colour = "Normal")) +
+      theme_bw() + theme(text = element_text(size=15)) +
+      scale_colour_manual("Groups", values = c("darkorange1", "light blue", "royal blue"))
> p
>
```



Двухвыборочный t-test



Необходимо проверить условия:

- Нормальность
- Независимость
- Равенство дисперсий

Равенство дисперсий по критерию Фишера:

$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{\max(s_1^2,\ s_2^2)}{\min(s_1^2,\ s_2^2)}$$

Пример:

R Console

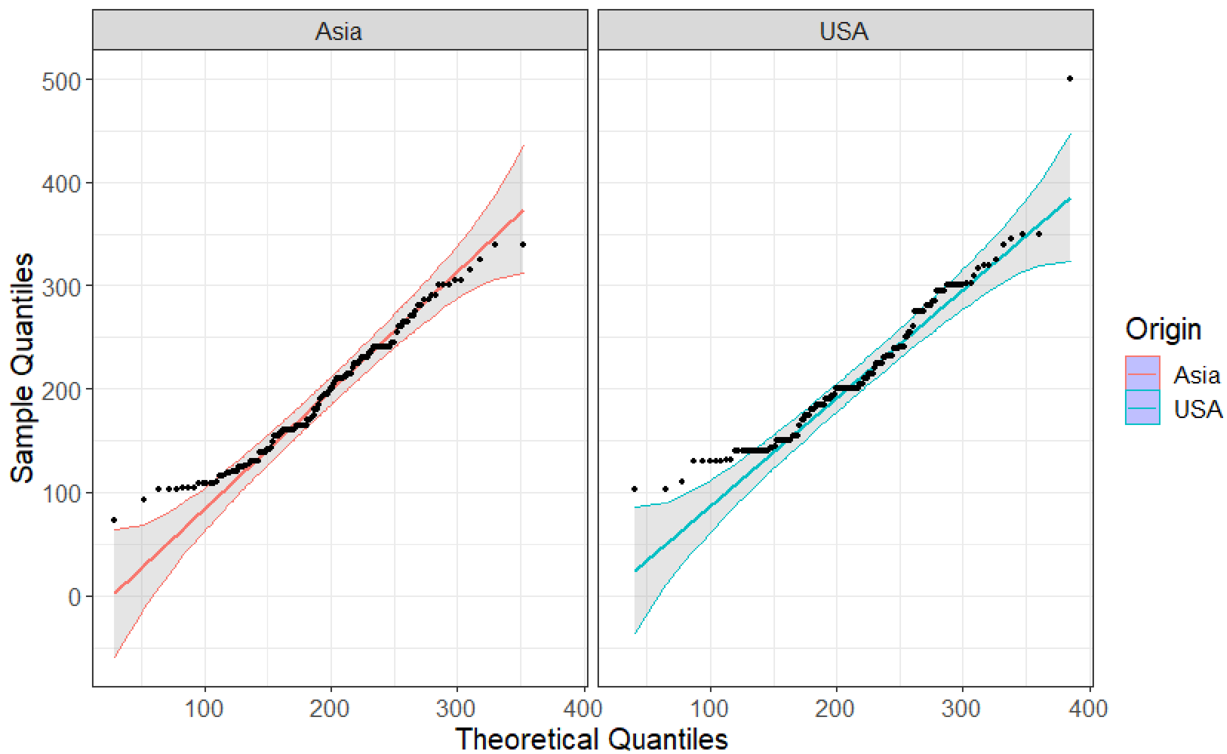
```
> sample_USA_Asia <- subset(cars, (Origin == "USA") | (Origin == "Asia"))
> var.test(Horsepower ~ Origin, data=sample_USA_Asia, alternative = "two.sided")

        F test to compare two variances

data:  Horsepower by Origin
F = 0.86801, num df = 157, denom df = 146, p-value = 0.3835
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6297812 1.1940994
sample estimates:
ratio of variances
         0.8680076
```

Выполнены ли условия применимости (уровень значимости 0.05)?

Двухвыборочный t-test

R Console

```
> ggplot(data = sample_USA_Asia, aes(sample = Horsepower, color = Origin)) +
+   stat_qq_band(alpha=0.25, conf=0.95) +
+   stat_qq_line() +
+   stat_qq_point(col="black", size = 1) +
+   labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
+   theme_bw() + theme(text = element_text(size=15)) +
+   facet_wrap(~Origin)
```



В случае неравных дисперсий

- Аппроксимация
  - Стандартная ошибка:

$$\mathrm{SE}_u = \left(\frac{s_1^2}{\sum_{i=1}^{n_1^\star} f_{1i} w_{1i}} + \frac{s_2^2}{\sum_{i=1}^{n_2^\star} f_{2i} w_{2i}} \right)^{\frac{1}{2}}$$

«число» наблюдений с учетом «частот» и «весов»

  - «Число» степеней свободы:

$$\mathrm{df}_u = \frac{\mathrm{SE}_u^4}{\frac{s_1^4}{(n_1-1)\left(\sum_{i=1}^{n_1^\star} f_{1i} w_{1i}\right)^2} + \frac{s_2^4}{(n_2-1)\left(\sum_{i=1}^{n_2^\star} f_{2i} w_{2i}\right)^2}}$$

  - Аппроксимированная t-статистика:

$$t_u = \frac{\bar{y}_d - \mu_0}{\mathrm{SE}_u}$$

  - Расчет p-value:

$$p\text{-value} = \begin{cases} P\left(t_u^2 > F_{1-\alpha,1,\mathrm{df}_u}\right) \ , & \text{2-sided} \\ P\left(t_u < t_{\alpha,\mathrm{df}_u}\right) \ , & \text{lower 1-sided} \\ P\left(t_u > t_{1-\alpha,\mathrm{df}_u}\right) \ , & \text{upper 1-sided} \end{cases}$$

Двухвыборочный t-test

```
> by(sample_USA_Asia$Horsepower, sample_USA_Asia$Origin, describe)
sample_USA_Asia$Origin: Asia
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 158 190.7 59.39  187.5  187.77 67.46  73 340   267 0.35    -0.65 4.73
------------------------------------------------------------
sample_USA_Asia$Origin: USA
   vars   n   mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 147 212.82 63.75    200  208.53 66.72 103 500   397 0.88     1.49 5.26
```

Различаются ли средние с уровнем значимости 0.05?

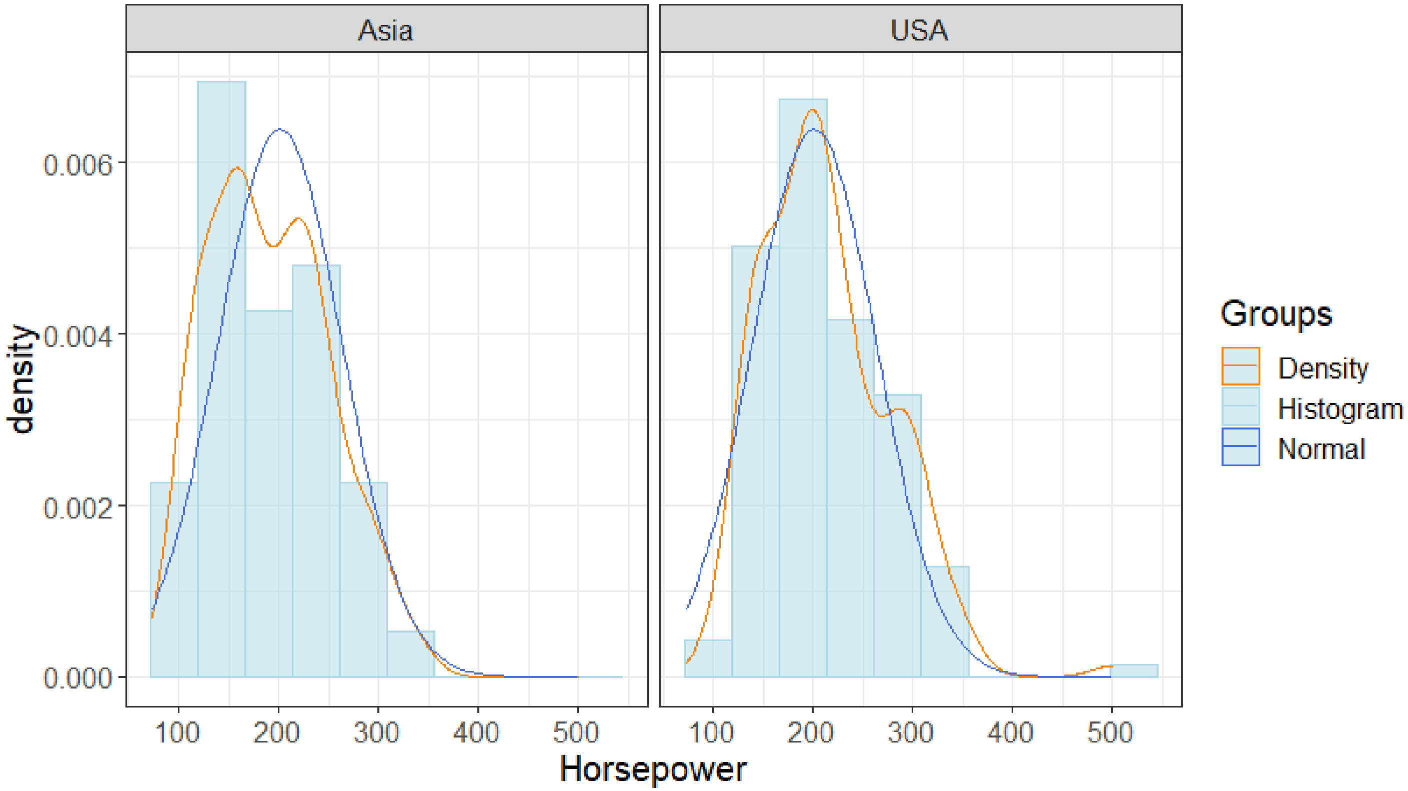
```
> t.test(Horsepower ~ Origin,
+ data=sample_USA_Asia, var.equal = FALSE)

        Welch Two Sample t-test

data:  Horsepower by Origin
t = -3.1292, df = 296.94, p-value = 0.001927
alternative hypothesis: true difference in means
95 percent confidence interval:
 -36.032364  -8.208831
sample estimates:
mean in group Asia  mean in group USA
          190.7025           212.8231
```

```
> t.test(Horsepower ~ Origin,
+ data=sample_USA_Asia, var.equal = TRUE)

        Two Sample t-test

data:  Horsepower by Origin
t = -3.1372, df = 303, p-value = 0.001873
alternative hypothesis: true difference in means
95 percent confidence interval:
 -35.995705  -8.245491
sample estimates:
mean in group Asia  mean in group USA
          190.7025           212.8231
```

Какой критерий использовать зависит от проверки на равенство дисперсий

Двухвыборочный t-test

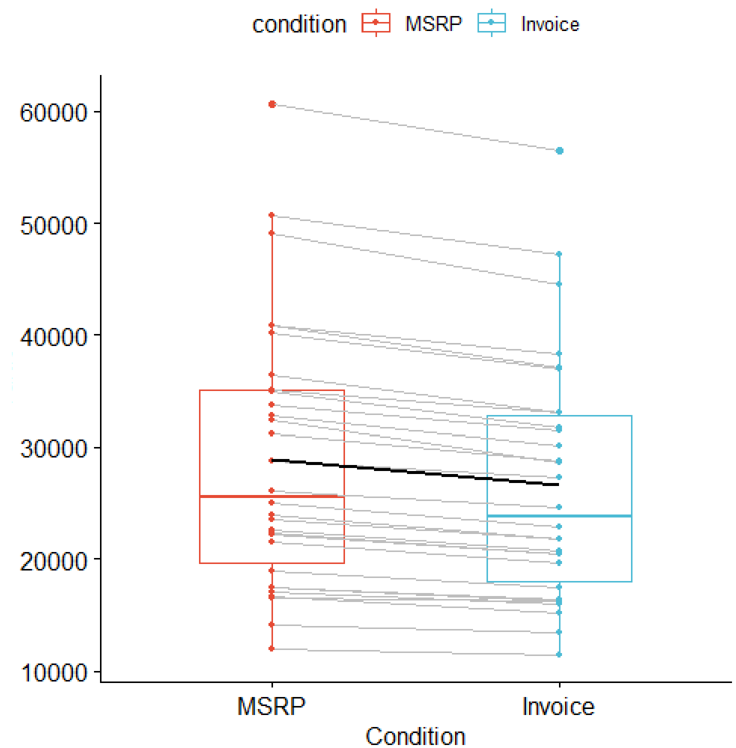
```
> ggplot(sample_USA_Asia, aes(x = Horsepower, group = Origin)) +
+ geom_histogram(aes(y = ..density.., x=Horsepower)) +
+ geom_density(aes(y = ..density.., x=Horsepower)) +
+ stat_function(fun = dnorm,
+               args = with(sample_USA_Asia, c(mean = mean(Horsepower), sd = sd(Horsepower)))) +
+ facet_wrap(~Origin)
```



Попарный t-test

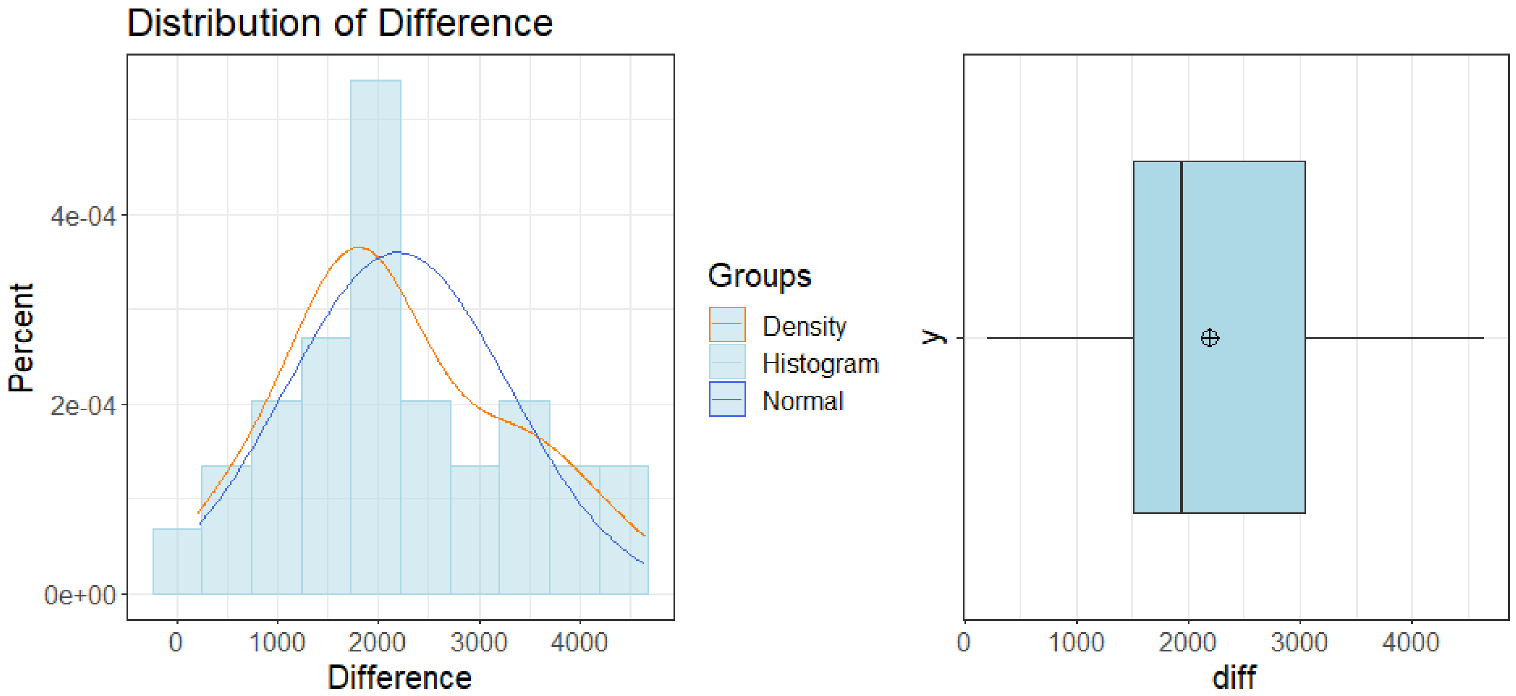
```
> sample_Wagon <- subset(cars, Type == "Wagon")
> sample_Wagon$diff <- sample_Wagon$MSRP - sample_Wagon$Invoice
> describe(sample_Wagon[,c("MSRP","Invoice","diff")])
        vars  n     mean       sd median  trimmed      mad   min   max range skew kurtosis      se
MSRP       1 30 28840.53 11834.00  25545 27592.46 12086.90 11905 60670 48765 0.77    -0.06 2160.58
Invoice    2 30 26645.63 10856.11  23721 25467.88 11012.75 11410 56474 45064 0.82     0.07 1982.05
diff       3 30  2194.90  1109.71   1940  2158.88   967.40   206  4644  4438 0.37    -0.67  202.60
```

```
> t.test(sample_Wagon$MSRP, sample_Wagon$Invoice, paired=TRUE)

        Paired t-test

data:  sample_Wagon$MSRP and sample_Wagon$Invoice
t = 10.833, df = 29, p-value = 1.041e-11
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 1780.529 2609.271
sample estimates:
mean difference
         2194.9
```




```
> ggpaired(sample_Wagon, cond1="MSRP", cond2="Invoice",
+          color = "condition", line.color = "gray") +
+     stat_summary(fun=mean, geom="line", size=1, aes(group = 1))
```

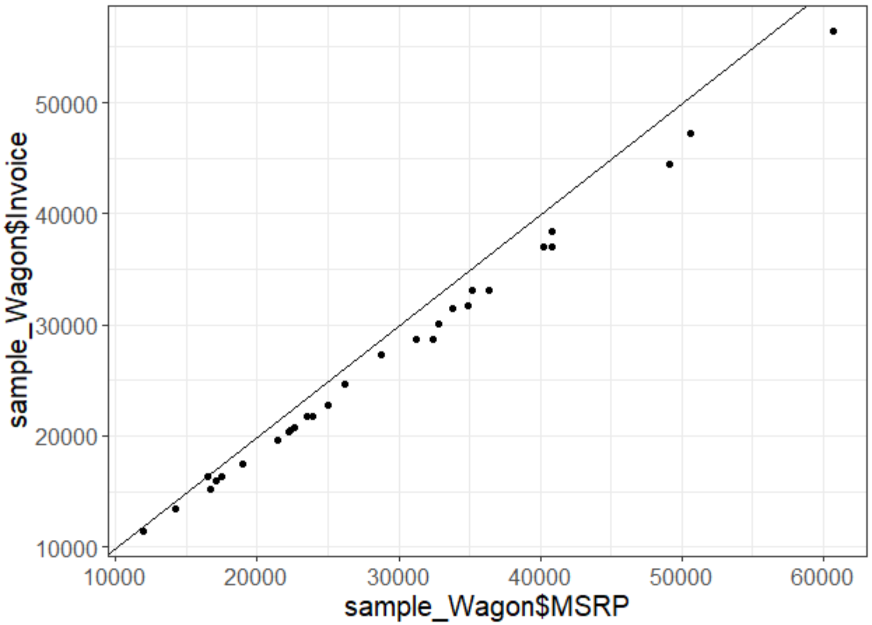
Попарный t-test

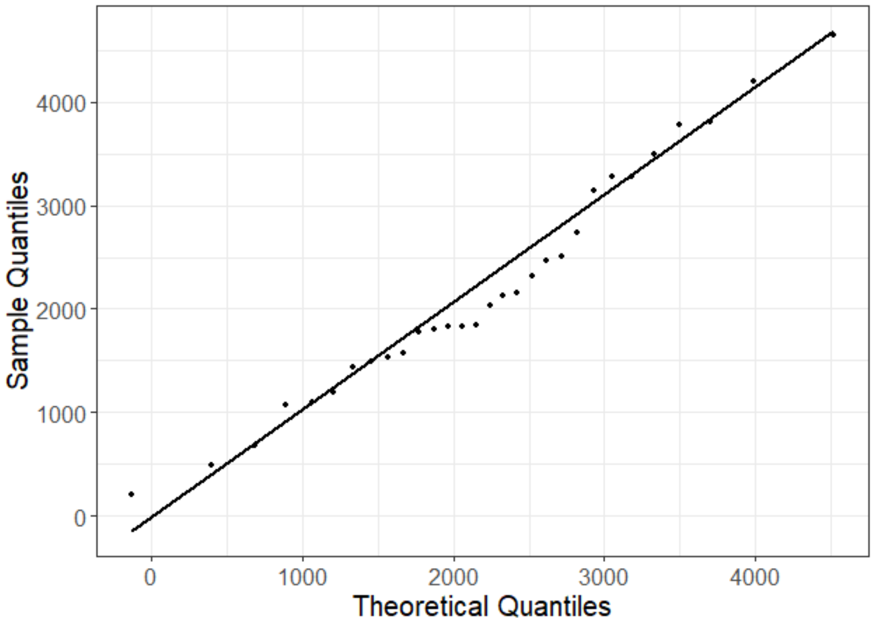



```
> dnorm_diff_pars = with(sample_Wagon, c(mean(diff), sd(diff)))
> p1 <- ggplot(sample_Wagon, aes(x = diff)) +
+       geom_histogram(aes(y = ..density..)) +
+       geom_density(aes(y = ..density..), alpha=.5) +
+       stat_function(fun = dnorm, args = dnorm_diff_pars)
>
> p2 <- ggplot(sample_Wagon, aes(x = diff, y = "")) +
+       geom_boxplot(fill = "light blue", outlier.alpha = 0.1) +
+       stat_summary(fun=mean, geom="point") +
+       theme_bw() + theme(text = element_text(size=15))
> p1 | p2
```

Попарный t-test

```
> ggplot(mapping = aes(x = sample_Wagon$MSRP, y = sample_Wagon$Invoice)) +
+   geom_point() +
+   geom_abline(aes(slope = 1, intercept = 0), linetype = 1) +
+   theme_bw() + theme(text = element_text(size=15))
```






```
> ggplot(data = sample_Wagon, mapping = aes(sample = diff)) +
+   stat_qq_line() +
+   stat_qq_point(col="black", size = 1) +
+   labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
+   theme_bw() + theme(text = element_text(size=15))
```

Непараметрический тест оценки «среднего» для двух выборок

- Wilcoxon rank-sum test –аналог двух выборочного t-test
  - Пусть дано две упорядоченные по возрастанию выборки $\{y_{11}, y_{12}, \dots, y_{1N1}\}$ и $\{y_{21}, y_{22}, \dots, y_{2N2}\}$
  - Обозначим ранги как $r_{ki} = rank(y_{ki})$ и суммы рангов групп как $R_k = \sum_{j=1}^{Nk} r_{kj}$
  - Базовая гипотеза как и в t-test – «нет разницы между R_1/n_1и R_2/n_2
  - Критерий: $$Z = \frac{|R_1 - \mu_{R1}| - 0.5}{\sigma_{R1}}$$
  -

$$\mu_{R1} = \left(\frac{n_1}{N}\right) \cdot \left(\frac{N\,(N+1)}{2}\right) = \frac{n_1 \cdot (N+1)}{2} \qquad \sigma^2_{R_1} = \frac{n_1 \cdot n_2}{12}(N+1)$$

  - Проверка есть точная и приближенная по нормальному распределению:

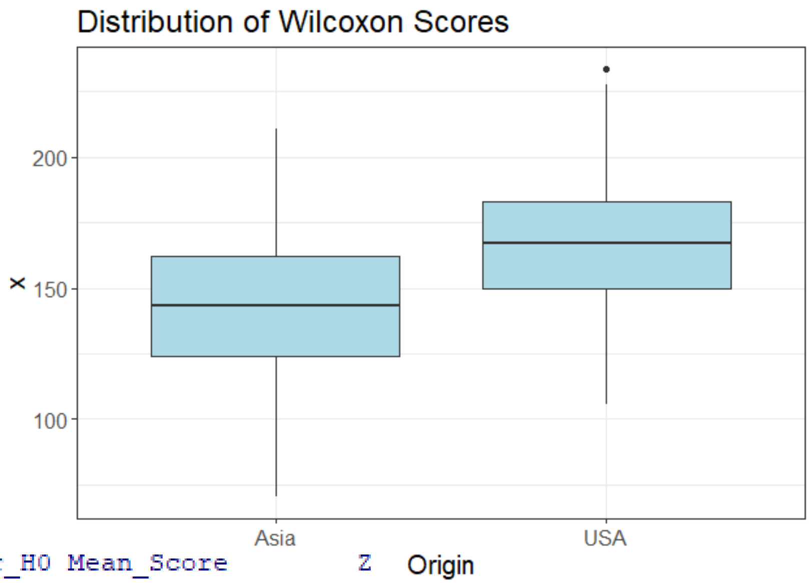
$$\text{reject } H_0 \text{ if } |Z| > Z_{\alpha/2}$$

Пример непараметрического сравнения средних

```
> sample_USA_Asia$rank <- rank(sample_USA_Asia$Horsepower)
> sample_USA <- subset(sample_USA_Asia, Origin == "USA")
> sample_Asia <- subset(sample_USA_Asia, Origin == "Asia")
```

```
> n1 <- nrow(sample_USA)
> n2 <- nrow(sample_Asia)
> N <- n1 + n2
>
> rank_1 <- sum(sample_USA$rank)
> rank_2 <- sum(sample_Asia$rank)
> mu_r1 <- n1*(N+1)/2
> mu_r2 <- n2*(N+1)/2
> sigma_H0 <- sqrt(n1*n2*(N+1)/12)
>
> mean_score1 <- rank_1 / n1
> mean_score2 <- rank_2 / n2
> z <- (abs(rank_1 - mu_r1) - 0.5) / sigma_H0
>
> dfr = data.frame(
+     Origin = c("USA", "Asia"),
+     N = c(n1, n2),
+     Sum_of_Scores = c(rank_1, rank_2),
+     Expect_Under_H0 = c(mu_r1, mu_r2),
+     Std_Under_H0 = rep(sigma_H0, 2),
+     Mean_Score = c(mean_score1, mean_score2),
+     Z = rep(z, 2)
+ )
> print(dfr)
  Origin   N Sum_of_Scores Expect_Under_H0 Std_Under_H0 Mean_Score        Z
1    USA 147       24618.5           22491     769.5863   167.4728 2.763823
2   Asia 158       22046.5           24174     769.5863   139.5348 2.763823
```

```
> wilcox.test(Horsepower ~ Origin, data = sample_USA_Asia, correct = TRUE)

        Wilcoxon rank sum test with continuity correction

data:  Horsepower by Origin
W = 9485.5, p-value = 0.005698
alternative hypothesis: true location shift is not equal to 0
```



Непараметрический тест оценки «среднего» для одной выборки

- Wilcoxon signed-rank test –аналог одновыборочного t-test
  - Пусть y_i разности (ненулевые), упорядоченные по возрастанию $|y_i|$ и R_i - ранг.
  - Пусть $R^{(+)}$ - множество рангов положительных y_i, а $R^{(-)}$ отрицательных
  - Базовая гипотеза как и в t-test – «нет разницы между $R^{(+)}$и $R^{(-)}$
  - $S = \left(R^{(+)} - R^{(-)}\right)/2$ и $V = (\mathrm{n}(\mathrm{n}+1)(2\mathrm{n}+1))\ /\ 24$
  - Критерий: $$T = \frac{S \cdot \sqrt{n-1}}{\sqrt{n \cdot V - S^2}}$$
  - Есть точный тест или аппроксимированный по распределению стьюдента:

$$\text{reject } H_0 \text{ if } |T| > t_{\alpha/2,n-1}$$

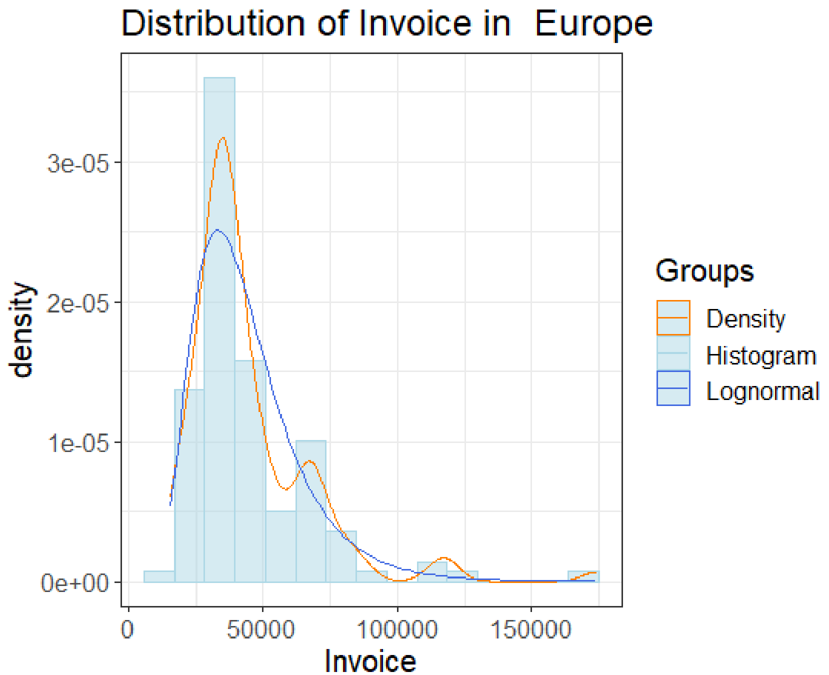
Оценка «среднего»

- Проверка гипотезы о заданном параметре положения (мат. ожидании)

```
> sample_Europe <- subset(cars, Origin == "Europe")
```

```
> t.test(sample_Europe$Invoice, mu = 45000)

        One Sample t-test

data:  sample_Europe$Invoice
t = -0.29067, df = 122, p-value = 0.7718
alternative hypothesis: true mean is not equal to 45000
95 percent confidence interval:
 40275.36 48514.80
sample estimates:
mean of x
 44395.08

> wilcox.test(sample_Europe$Invoice, mu = 45000)

        Wilcoxon signed rank test with continuity correction

data:  sample_Europe$Invoice
V = 2959, p-value = 0.03122
alternative hypothesis: true location is not equal to 45000

> sign_test(Invoice ~ 1, data = sample_Europe, mu = 45000)
# A tibble: 1 × 7
  .y.     group1 group2         n statistic    df        p
* <chr>   <chr>  <chr>      <int>     <dbl> <dbl>    <dbl>
1 Invoice 1      null model   123        39   123 0.000061
```

```
> print_log_distr(sample_Europe)
```



Проверка соответствия эмпирического распределения теоретическому

- Колмогоров-Смирнов:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

```
ks.test(x, y, …, alternative = c("two.sided",
"less", "greater"), exact = NULL, tol=1e-8,
simulate.p.value=FALSE, B=2000)
```

- Андерсон-Дарлинг:

$$-n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln(F(x_i,\theta)) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i,\theta))\right\}$$

```
ad.test(x)
```

- Крамер — Мизес:

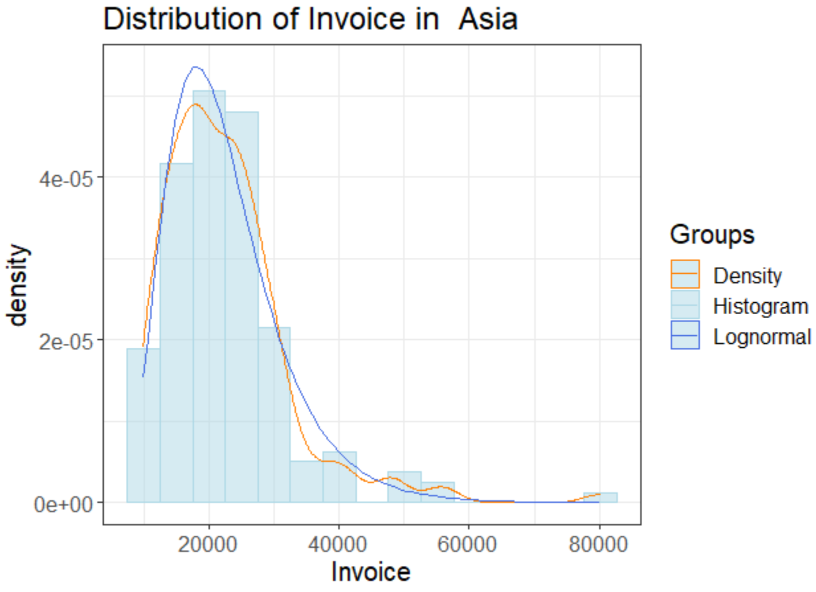
$$\frac{1}{12n} + \sum_{i=1}^{n}\left(F(x_i,\theta) - \frac{2i-1}{2n}\right)^2$$

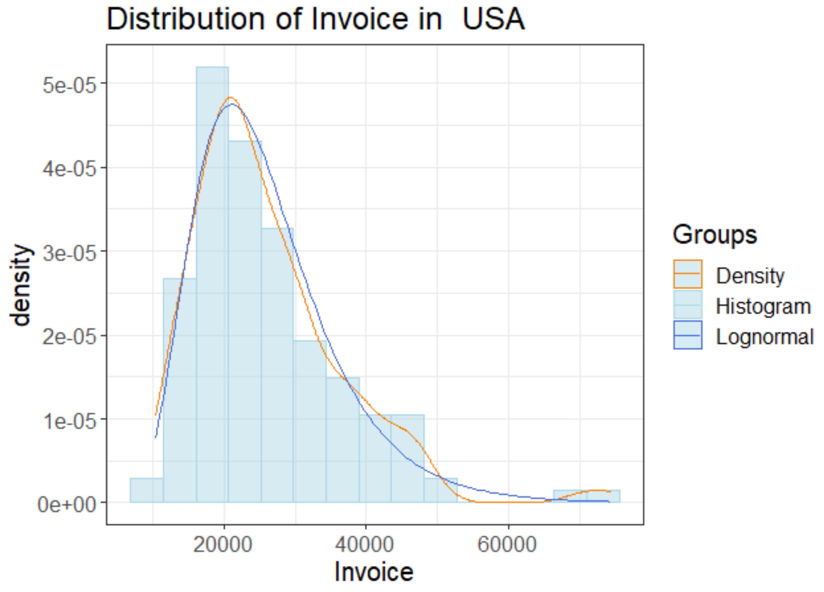
```
cvm.test(x, null = "punif", ...,
estimated=FALSE, nullname)
```

Описание переменной с помощью Log-Normal

$$p(x) = \begin{cases} \frac{h\nu}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \leq \theta \end{cases}$$

```
> get_dln_params <- function(x) c(mean(log(x)), sd(log(x)))
>
> print_log_distr <- function(df){
+     params <- get_dln_params(df$Invoice)
+     ggplot(df, aes(x = Invoice)) +
+      geom_histogram(aes(y = ..density.., colour = "Histogram"), fill="light blue", bins=15, alpha=.5) +
+      geom_density(aes(y = ..density.., colour = "Density"), alpha=.5) +
+      stat_function(fun = dlnorm, args = params, aes(colour = "Lognormal")) +
+      theme_bw() + theme(text = element_text(size=15)) +
+      ggtitle(paste("Distribution of Invoice in ", df$Origin[1])) +
+      scale_colour_manual("Groups", values = c("darkorange1", "light blue", "royal blue"))
+ }
>
> print_log_distr(sample_Asia)
> print_log_distr(sample_USA)
```





Описание переменной с помощью Log-Normal

$$p(x) = \begin{cases} \frac{h\nu}{\sigma\sqrt{2\pi}(x-\theta)} \exp\left(-\frac{(\log(x-\theta)-\zeta)^2}{2\sigma^2}\right) & \text{for } x > \theta \\ 0 & \text{for } x \le \theta \end{cases}$$

```
> feature <- sample_USA_Asia$Invoice
```

```
> ks.test(feature, "pnorm", mean(feature), sd(feature))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  feature
D = 0.11039, p-value = 0.001183
alternative hypothesis: two-sided
```

```
> cvm.test(feature)

        Cramer-von Mises normality test

data:  feature
W = 1.1918, p-value = 7.37e-10
```

```
> ks.test(feature, "plnorm", mean(log(feature)), sd(log(feature)))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  feature
D = 0.029459, p-value = 0.9539
alternative hypothesis: two-sided
```

```
> cvm.test(log(feature))

        Cramer-von Mises normality test

data:  log(feature)
W = 0.045941, p-value = 0.5724
```

```
> ad.test(feature)

        Anderson-Darling normality test

data:  feature
A = 7.336, p-value < 2.2e-16
```

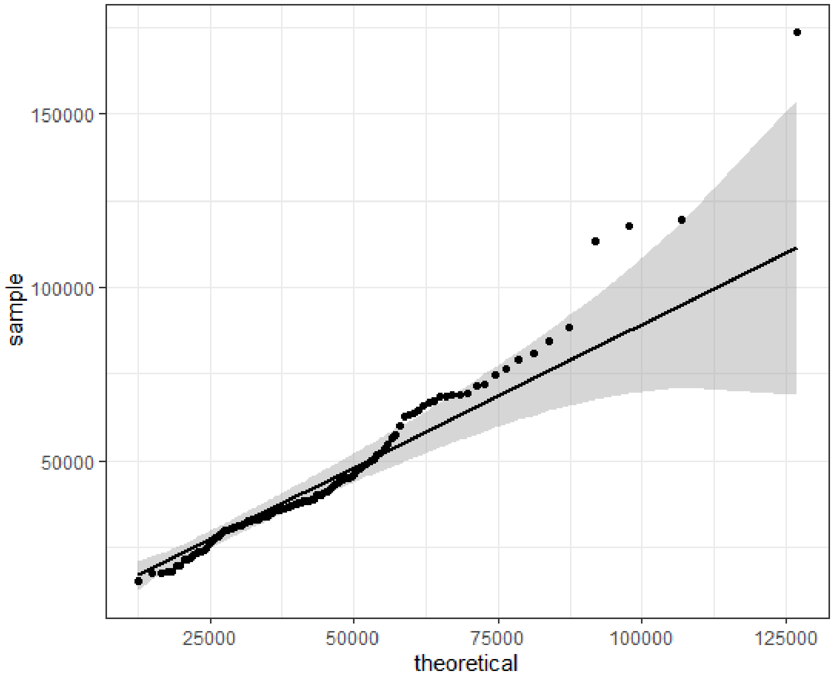
```
> ad.test(log(feature))

        Anderson-Darling normality test

data:  log(feature)
A = 0.38743, p-value = 0.3858
```

Анализ распределения переменной

- Проверка соответсвия квантилей и процентилей распределений

```
> sample_Europe <- subset(cars, Origin == "Europe")
```

```
> params <- get_dln_params(sample_Europe$Invoice)
> ggplot(sample_Europe, aes(sample = Invoice)) +
+ geom_qq_band(distribution = "lnorm", dparams = params, alpha = 0.4) +
+        stat_qq_line(distribution = "lnorm", dparams = params) +
+        stat_qq_point(distribution = "lnorm", dparams = params) +
+        theme_bw()
```



Рассматриваемые модели

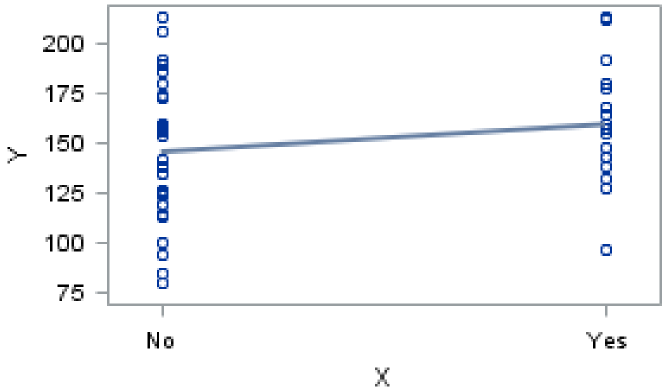
Предиктор / Отклик	**Категориальный**	**Непрерывный**	**Непрерывный и категориальный**
Непрерывный	Дисперсионный анализ (ANOVA)	Регрессия наименьших квадратов (OLS Regression)	Ковариационный анализ (ANCOVA)
Категориальный	Логистическая регрессия	Логистическая регрессия	Логистическая регрессия

Линейные модели

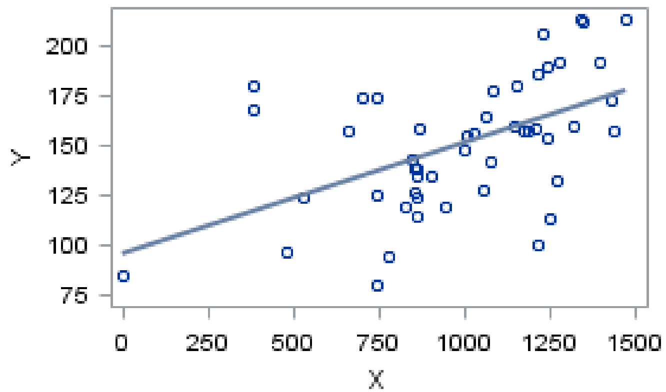
– Общая линейная регрессия

$$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k + \varepsilon$$

- Дисперсионный анализ (ANOVA)

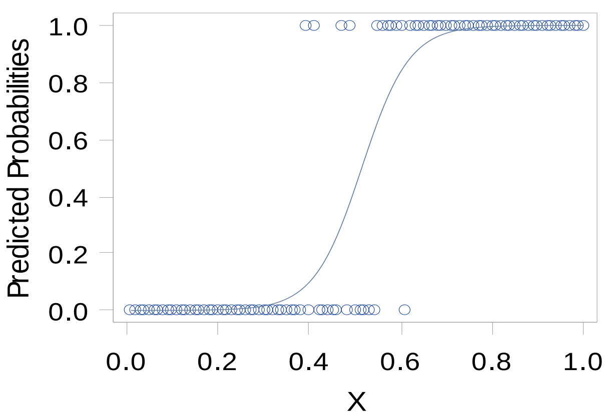


- Регрессия



– Логистическая регрессия

$$\mathrm{logit}(Y) = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k$$



Описательные и прогнозные модели

Описательные модели

$$\hat{Y}_i = \underline{\hat{\beta}}_0 + \underline{\hat{\beta}}_1 X_{1i} \ldots + \underline{\hat{\beta}}_k X_{ki}$$

- Как связаны X и Y?
- Интерпретируемость
- Небольшие выборки
- Мало переменных
- Оценка на основе *p*-values и доверительных интервалов

Прогнозные модели

$$\underline{\hat{Y}}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} \ldots + \hat{\beta}_k X_{ki}$$

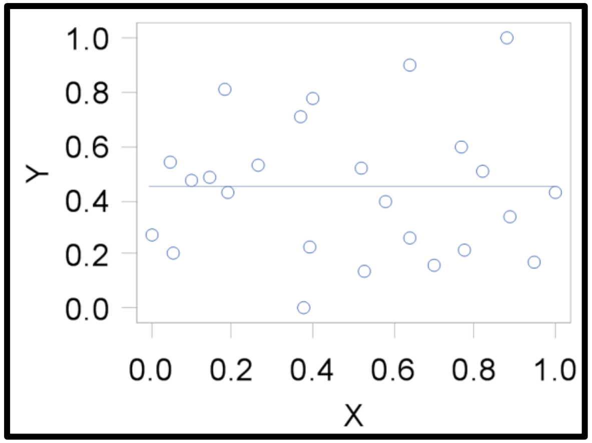
- Если знаем X_i, то прогнозируем Y_i
- Большие выборки
- Много переменных
- Оценки на валидационных и тестовых наборах

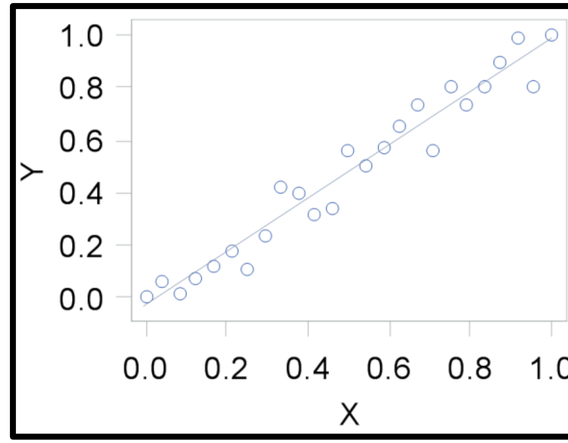
Разные зависимости с непрерывным откликом

Нет зависисмости

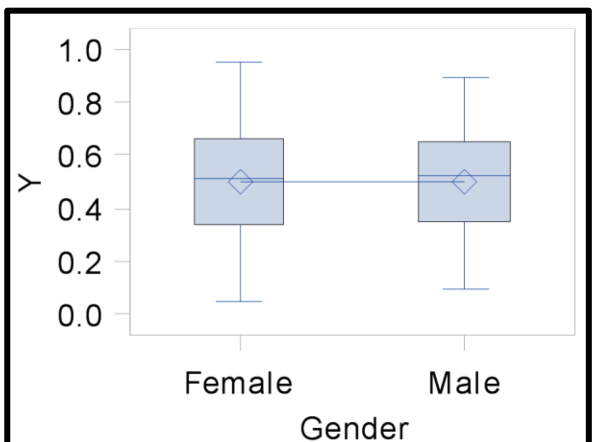
Есть зависисмость

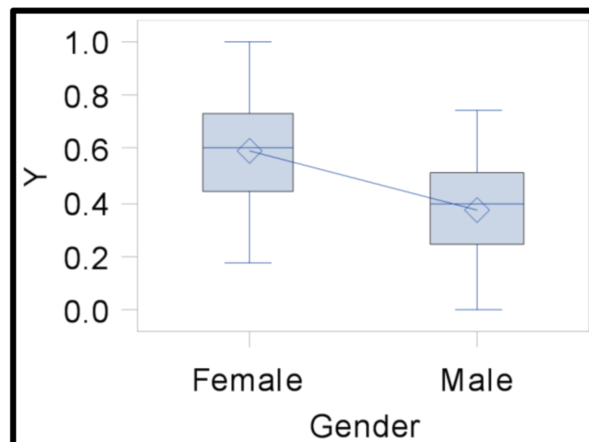
Непрервыный X





Категориальный X





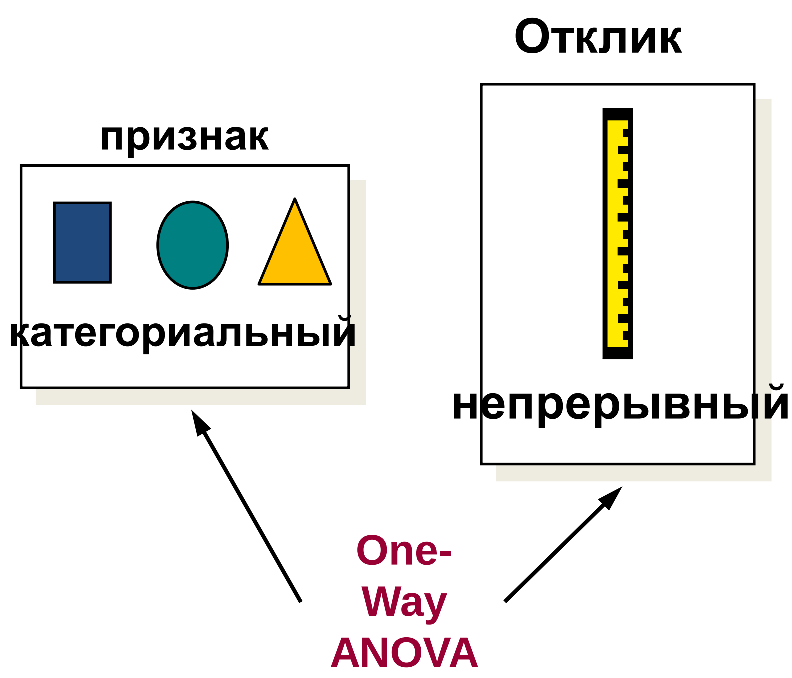
Рассматриваемые модели

Предиктор / Отклик	Категориальный	Непрерывный	Непрерывный и категориальный
Непрерывный	Дисперсионный анализ (ANOVA)	Регрессия наименьших квадратов (OLS Regression)	Ковариационный анализ (ANCOVA)
Категориальный	Логистическая регрессия	Логистическая регрессия	Логистическая регрессия

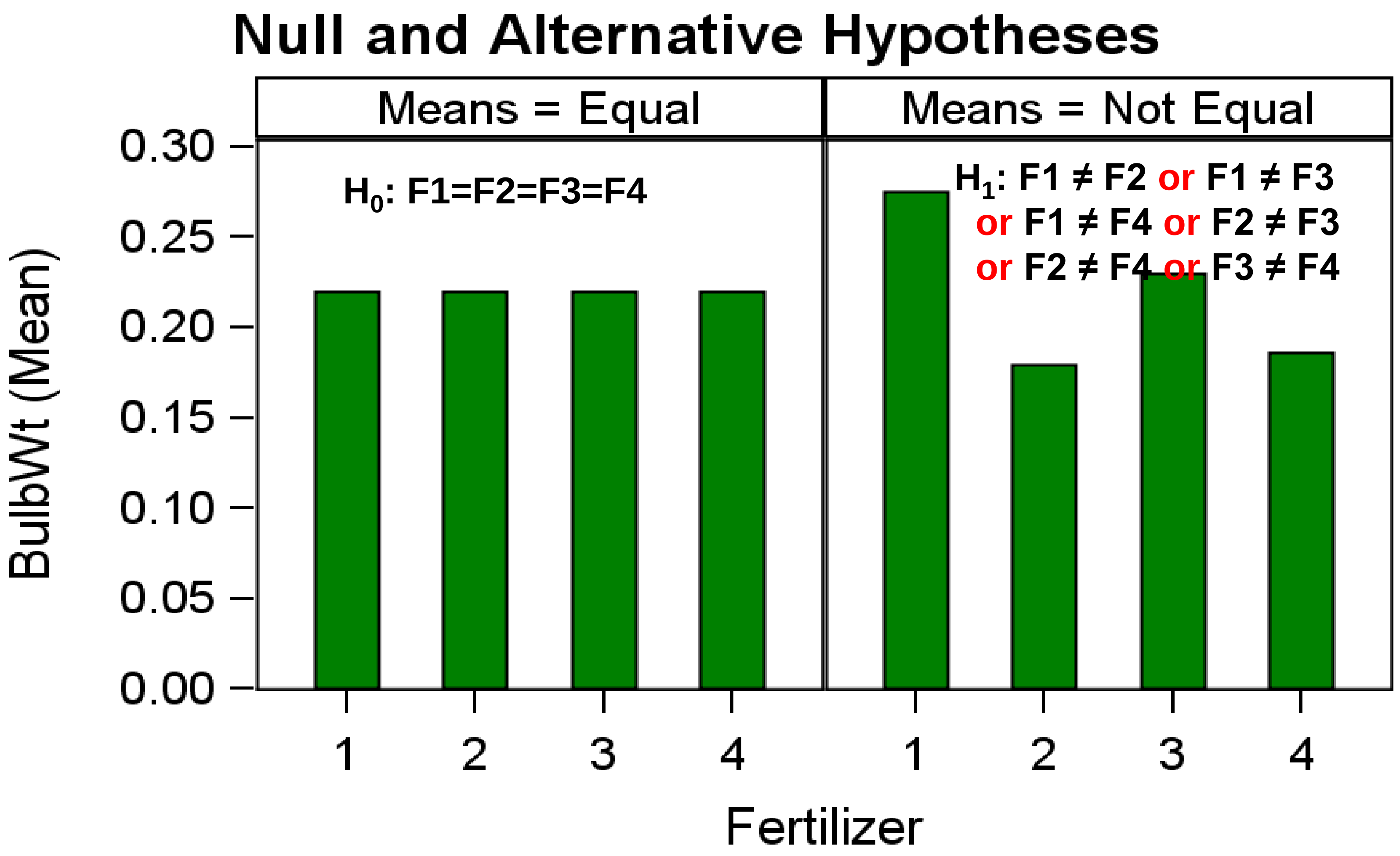
Дисперсионный анализ

- Основной вопрос:
  - различаются ли выборочные средние в группах?
  - поможет ли информация о принадлежности группе предсказать непрерывный отклик?
- Примеры задач:

- Действительно ли применение данного ***лекарства*** влияет на **артериальное давление**?
- Зависит ли **аварийность** от ***цвета*** автомобиля?
- Действительно ли в разных ***регионах*** разная **продолжительность жизни**?



Основная гипотеза дисперсионного анализа

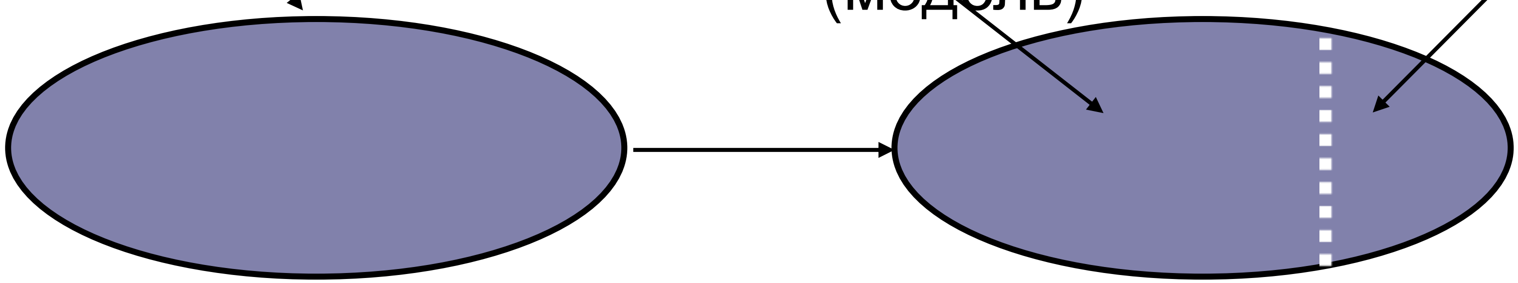


Представление общей дисперсии

Общая вариативность

Межгрупповая вариативность (модель)

Внутригрупповая вариативность (ошибка)



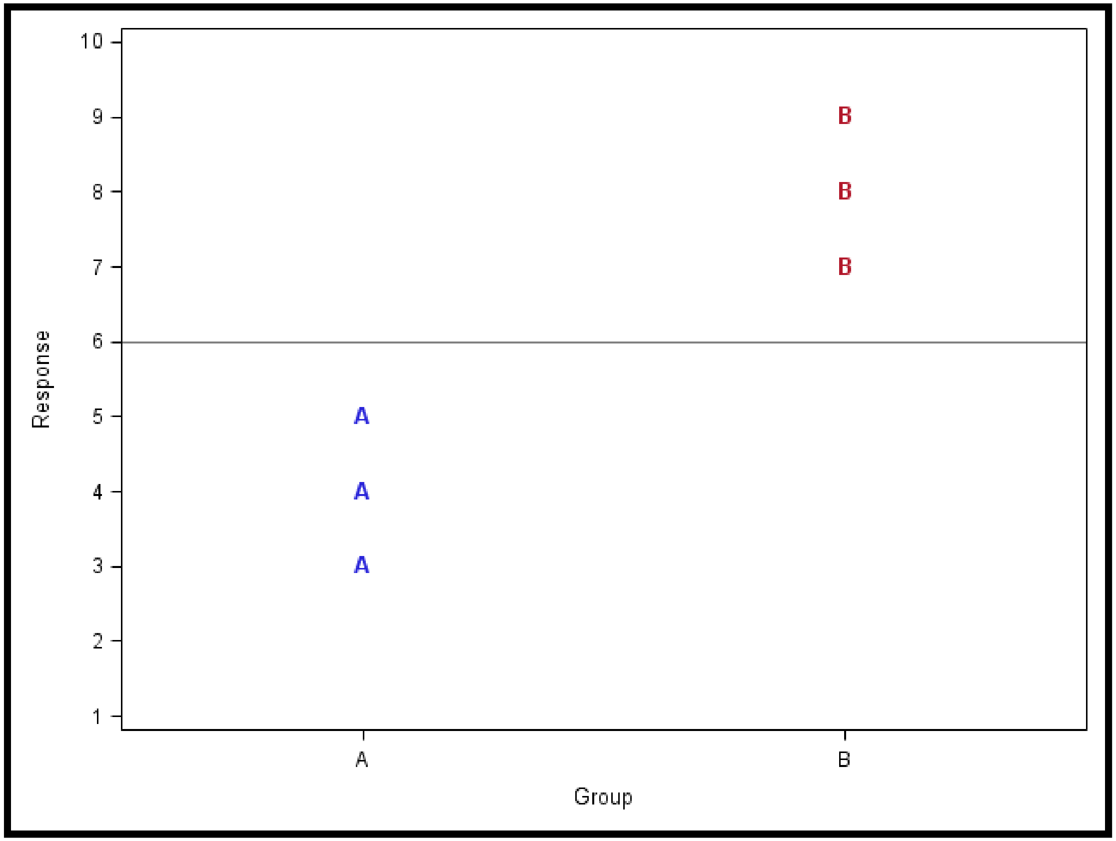
$$\mathbf{SS_T} = \mathbf{SS_M + SS_E}$$

$$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{B}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{..}\right)^2$$

$$\mathrm{SS}_{\mathrm{within}} = \sum_{i=1}^{B}\mathrm{SS}_i = \sum_{i=1}^{B}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{i\,.}\right)^2$$

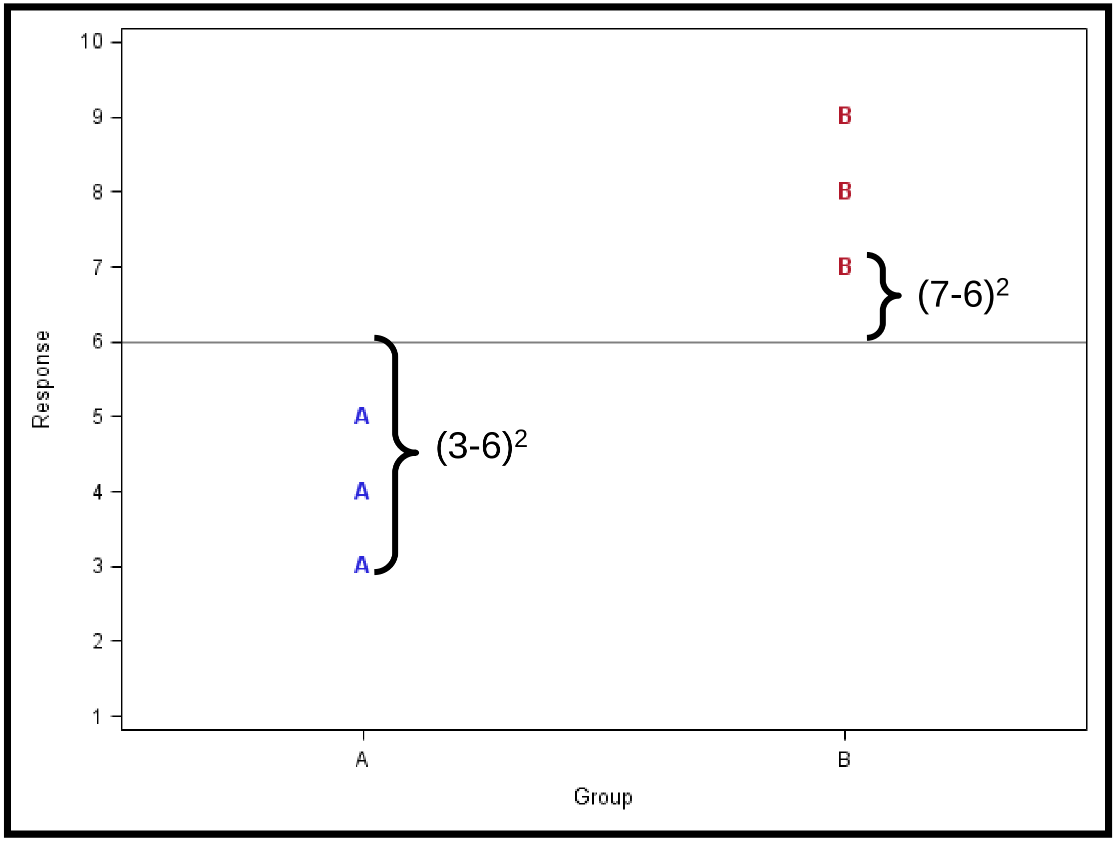
$$\mathrm{SS}_{\mathrm{between}} = \sum_{i=1}^{B} n_i\left(\overline{y}_{i\,.} - \overline{y}_{..}\right)^2$$

Пример: сумма квадратов



$$\bar{\bar{y}} = \frac{3+4+5+7+8+9}{6} = \mathbf{6}$$

Пример: сумма квадратов (общая)

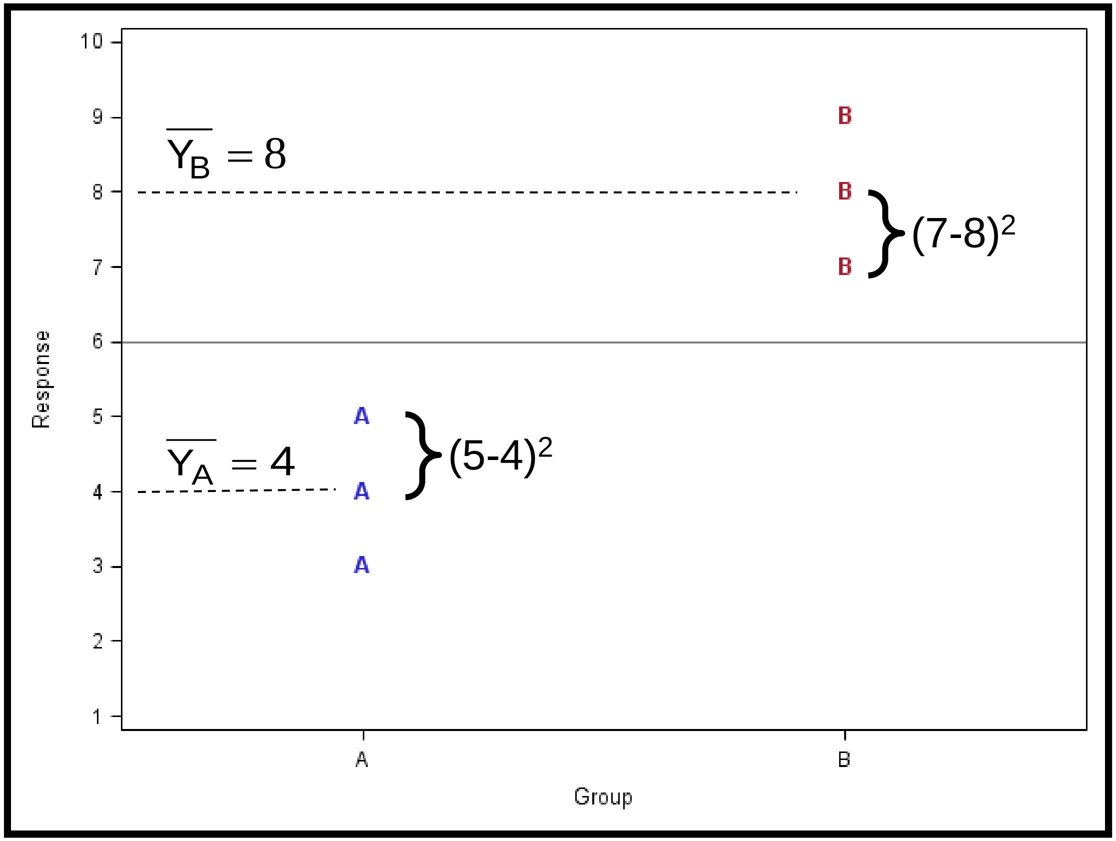


$SS_T =$

$(7-6)^2 +$
$(8-6)^2 +$
$(9-6)^2 +$

$(3-6)^2 +$
$(4-6)^2 +$
$(5-6)^2 +$
$= \mathbf{28}$

Пример: сумма квадратов (внутригрупповая, ошибка)



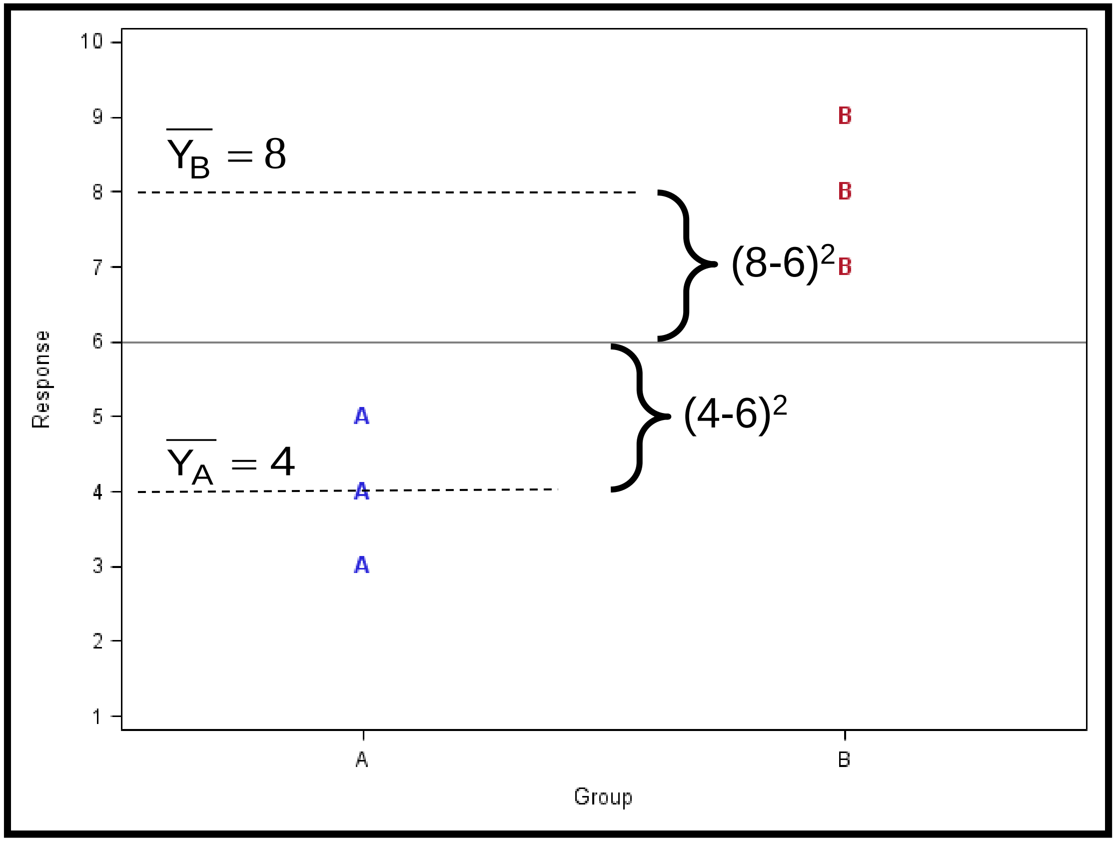
$SS_E =$

$(7-8)^2 +$
$(8-8)^2 +$
$(9-8)^2 +$

$(3-4)^2 +$
$(4-4)^2 +$
$(5-4)^2$

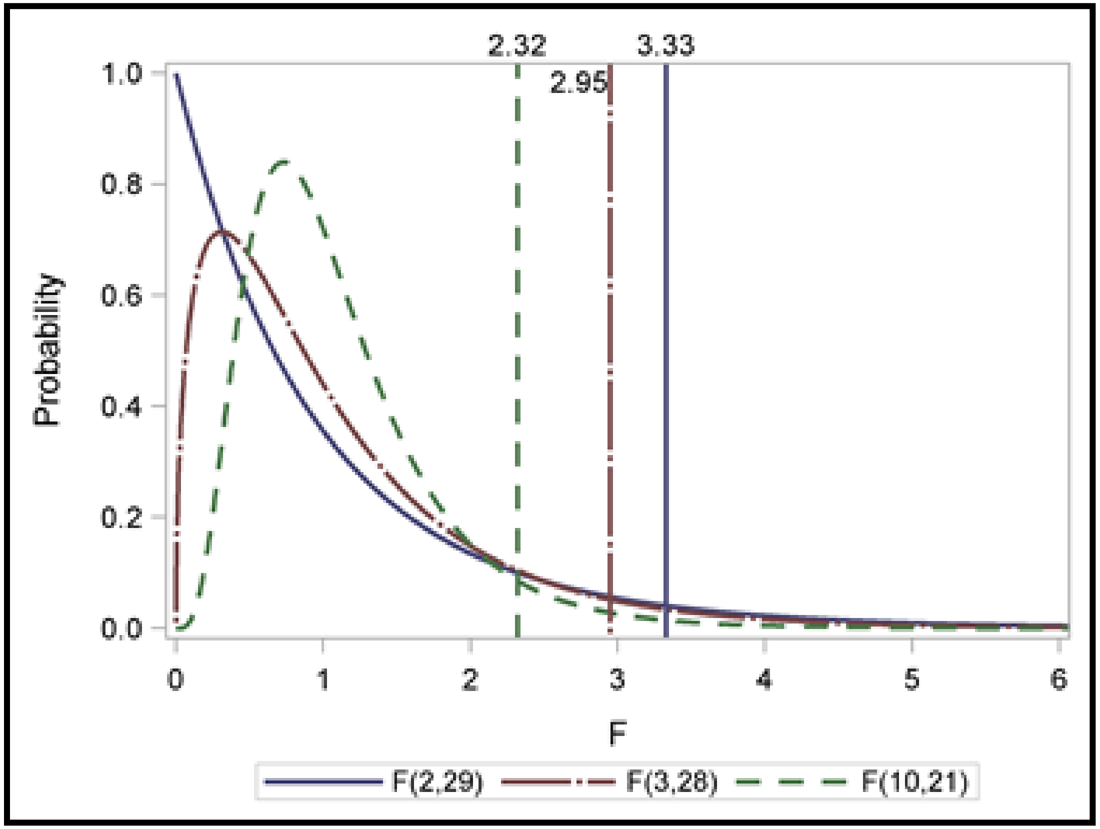
= 4

Пример: сумма квадратов (межгрупповая, модель)



$$SS_M = 3*(4-6)^2 + 3*(8-6)^2 = \mathbf{24}$$

Критерий Фишера



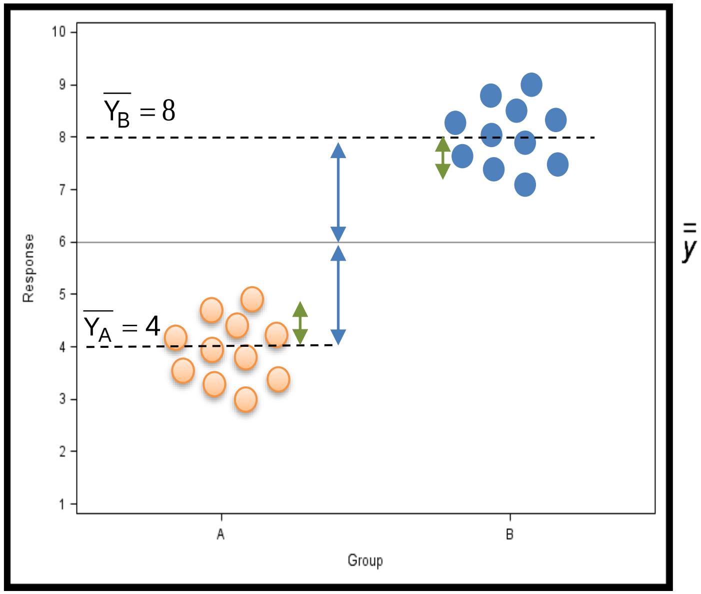
$$F(.,.) = \frac{MSM}{MSE} = \frac{\frac{SSM}{ModelDF}}{\frac{SSE}{ErrorDF}}$$

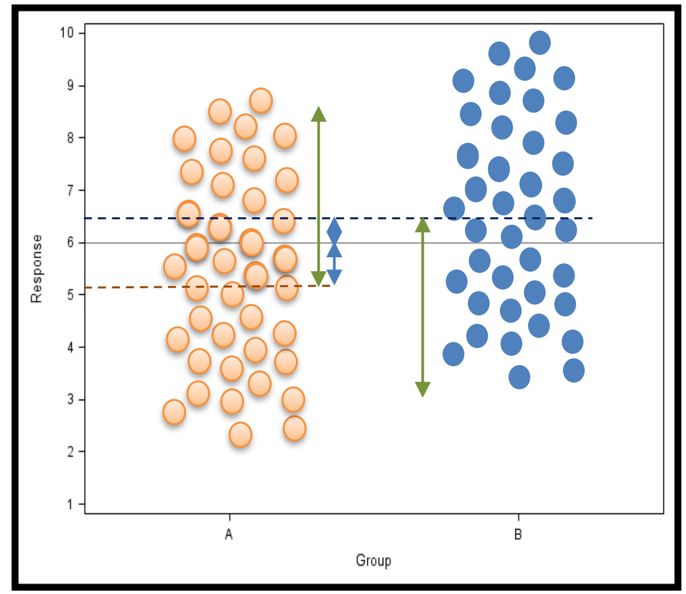
ModelDF = число групп -1

ErrorDF=Nobs -1 - (ModelDF)

$$F = \left(\frac{\mathrm{SS}_{\mathrm{between}}}{\mathrm{SS}_{\mathrm{within}}}\right)\left(\frac{n-B}{B-1}\right) \sim F_{B-1,\,n-B}$$

Коэффициент детерминации





F >> 1

F ~ 1

$$\mathbf{R^2 = SS_M / SS_T}$$

Пропорция вариации отклика, описываемая моделью (с заданным(и) предиктором(ами))

Модель ANOVA

Отклик = **База** (Среднее по выборке) + **Эффект** (разность среднего по группе и по выборке) + **Ошибка** (отличие от средн. по группе)

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

Основная процедура для ANOVA:

- Задаются категориальные переменные для идентификаторов групп
- Строится линейная регерссия с «бинарным» кодированием категориальных переменных
- Результат в терминах «групповых средних»

Предположения:

- независимость наблюдений,
- нормальность ошибки,
- равенство групповых дисперсий

Процедуры ANOVA

- Общий синтаксис lm:

```
lm(formula, data, subset, weights, na.action, method = "qr",
    model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
    singular.ok = TRUE, contrasts = NULL, offset, …)
```

- Общий синтаксис eov (наследуется от lm):

```
aov(formula, data = NULL, projections = FALSE, qr = TRUE,
    contrasts = NULL, …)
```

- Прогноз:

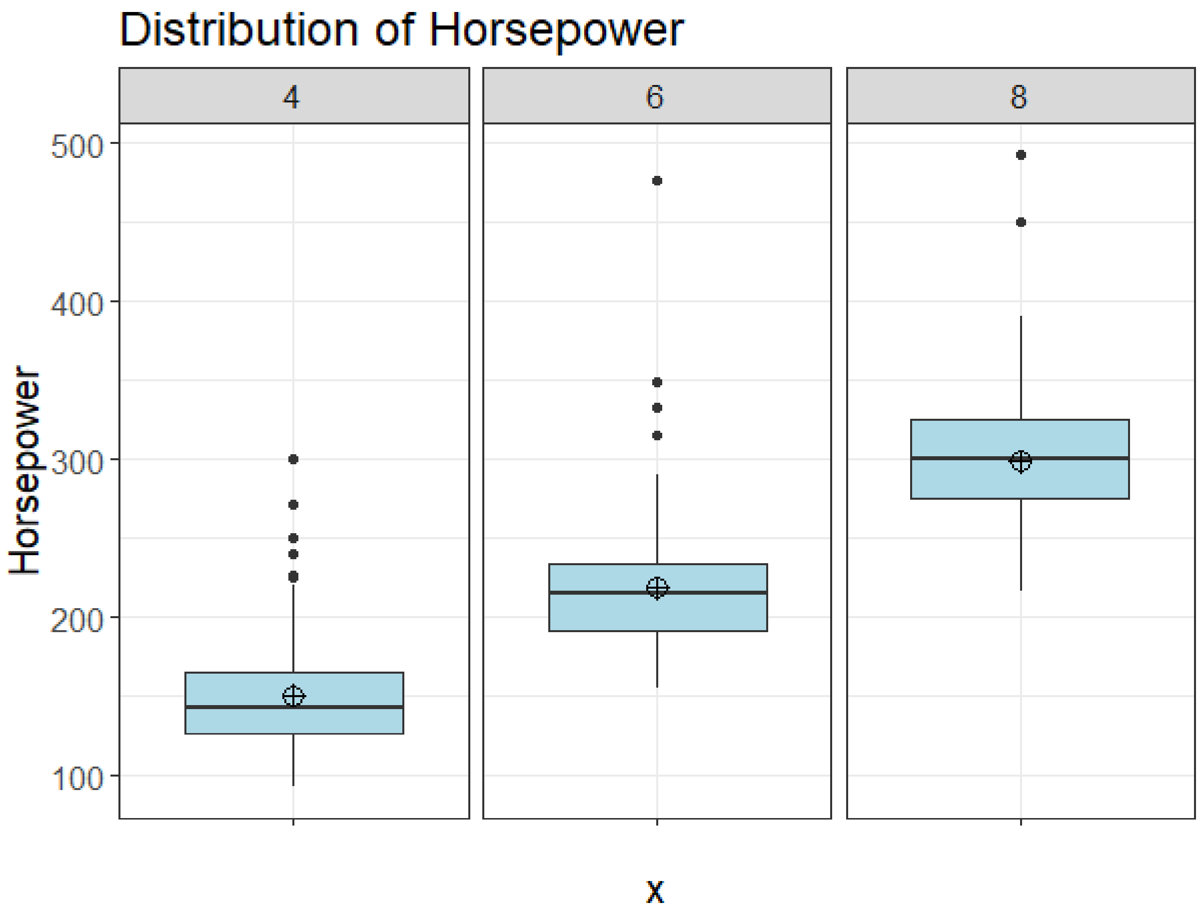
```
predict(object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
    interval = c("none", "confidence", "prediction"), level =
    0.95, type = c("response", "terms"), terms = NULL,
    na.action = na.pass, pred.var = res.var/weights, weights =
    1, vcov., ...)
```

- Остатки:

```
residuals(object, …)
```

Пример с проверкой на равенство дисперсий

```
> ggplot(sample_3_Cyl, aes(x = "", y = Horsepower, group = Cylinders)) +
+   geom_boxplot(fill = "light blue") +
+   stat_summary(fun=mean, geom="point", shape=10, size=3.5, color="black") +
+   ggtitle("Distribution of Horsepower") +
+   theme_bw() + theme(text = element_text(size=15)) +
+   facet_wrap(~Cylinders)
```



Пример с проверкой на равенство дисперсий

```
> sample_3_Cyl <- subset(cars, (Cylinders > 3) & (Cylinders != 5) & (Cylinders < 10))
> by(sample_3_Cyl$Invoice, sample_3_Cyl$Cylinders, describe)
sample_3_Cyl$Cylinders: 4
   vars   n     mean      sd  median trimmed    mad  min   max range skew kurtosis     se
X1    1 136 18352.67 6523.22 17380.5 17365.9 5180.2 9875 40883 31008 1.43     1.91 559.36
------------------------------------------------------------------------------------------------
sample_3_Cyl$Cylinders: 6
   vars   n     mean       sd median  trimmed     mad   min    max  range skew kurtosis      se
X1    1 190 29319.71 14347.43  26813 27283.89 7530.87 14978 173560 158582 5.95    52.78 1040.87
------------------------------------------------------------------------------------------------
sample_3_Cyl$Cylinders: 8
   vars  n    mean       sd median  trimmed      mad   min    max range skew kurtosis      se
X1    1 87 46550.4 17431.28  43556 44932.65 16456.86 19490 113388 93898 1.05      1.3 1868.83
```

```
> bartlett.test(Horsepower ~ Cylinders, data = sample_3_Cyl)

	Bartlett test of homogeneity of variances

data:  Horsepower by Cylinders
Bartlett's K-squared = 8.7563, df = 2, p-value = 0.01255
```

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^k (n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^k \left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)}$$

$$S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1) S_i^2$$

```
> aov_model <- aov(Horsepower ~ Cylinders, sample_3_Cyl)
> summary(aov_model)
             Df  Sum Sq Mean Sq F value Pr(>F)
Cylinders     2 1180295  590148   343.4 <2e-16 ***
Residuals   410  704595    1719
```

Пример с проверкой на равенство дисперсий

```
> lm_model <- lm(Horsepower ~ Cylinders, sample_3_Cyl)
> summary(lm_model)

Call:
lm(formula = Horsepower ~ Cylinders, data = sample_3_Cyl)

Residuals:
    Min      1Q  Median      3Q     Max
-81.943 -24.953  -3.953  16.057 258.047

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  150.515      3.555   42.34   <2e-16 ***
Cylinders6    68.438      4.656   14.70   <2e-16 ***
Cylinders8   148.428      5.691   26.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.46 on 410 degrees of freedom
Multiple R-squared:  0.6262,	Adjusted R-squared:  0.6244
F-statistic: 343.4 on 2 and 410 DF,  p-value: < 2.2e-16
```

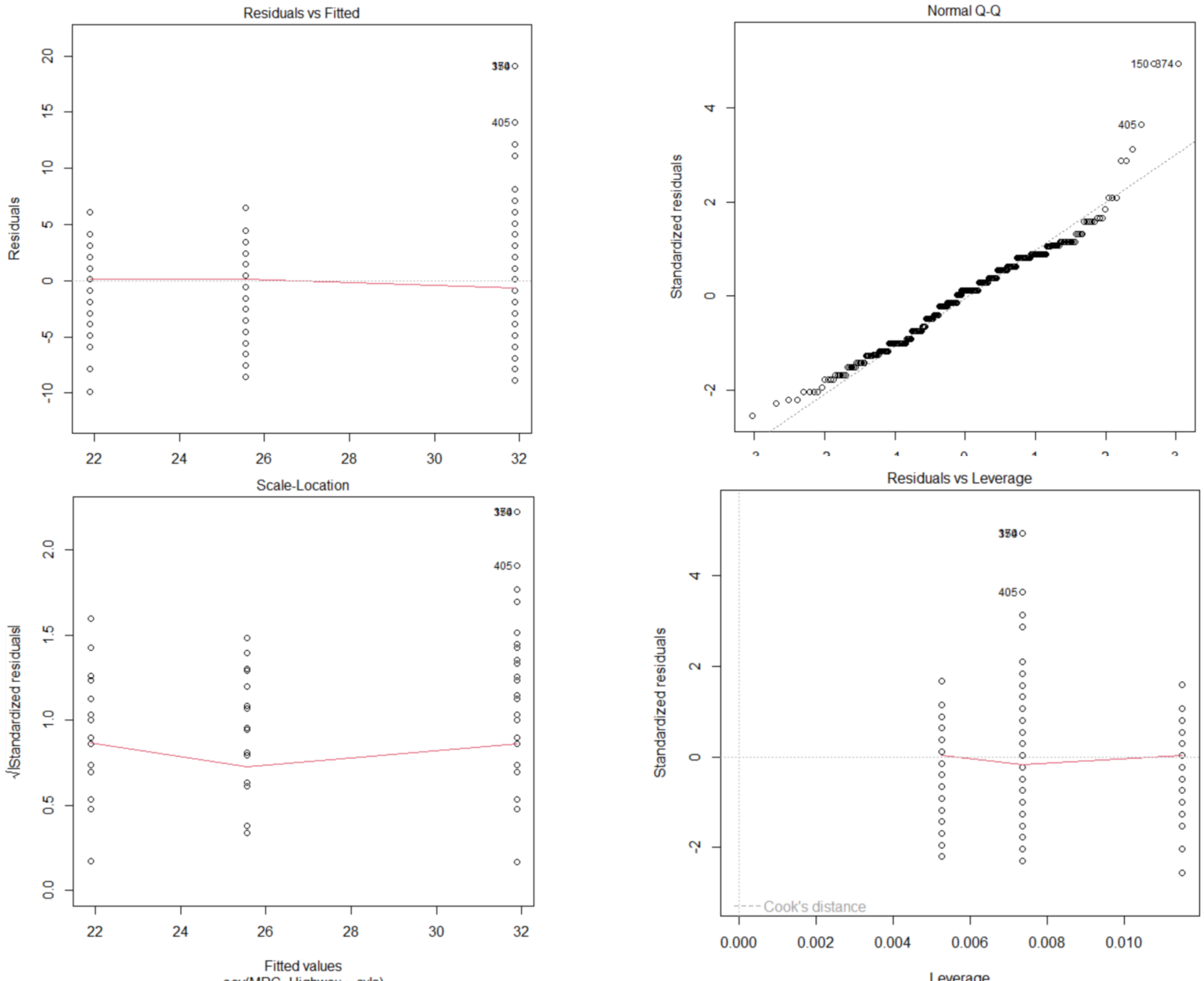
Прогнозы и остатки в ANOVA

На основе ANOVA и lm модели возможно прогнозирование для каждого наблюдения:

- Групповое среднее его группы
- Остатки – разность между реальным откликом и прогнозом
- Другие статистики

```
> pred <- predict(lm_model)
> resid <- residuals(lm_model)
> summary(pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  150.5   150.5   219.0   213.3   219.0   298.9 
> summary(resid)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-81.943 -24.953  -3.953   0.000  16.057 258.047 
> pred[1:5]
       1        2        3        4        5 
218.9526 150.5147 150.5147 218.9526 218.9526 
> resid[1:5]
        1         2         3         4         5 
46.047368 49.485294 49.485294 51.047368  6.047368 
```

Графики в процедурах для ANOVA



Множественные сравнения

Число групп	Число сравнений	Уровень ошибки всей серии (α=0.05)
2	1	.05
3	3	.14
4	6	.26
5	10	.40

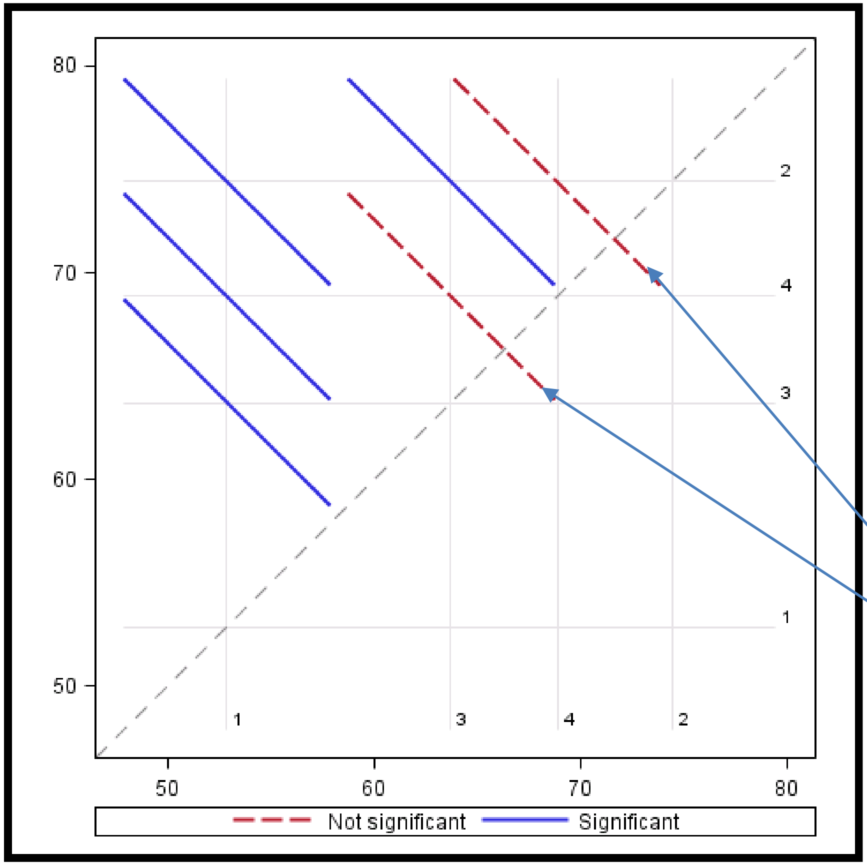
Comparisonwise (для каждого сравнения) Error Rate = α = 0.05

Для всей серии сравнений EER $\leq 1 - (1 - \alpha)^{nc}$, *nc*=число сравнений

Control Comparisonwise Error Rate → **Pairwise t-tests**

Control Experimentwise Error Rate → **Compare All Pairs Tukey**; **Compare to Control Dunnett**

Diffograms



- По горизонтали и вертикали – группы
- На пересечении (серые линии) оценка разброса «разности» средних в соответсвующих двух группах (по сути длина линии – доверительный интервал для попарной разности)
- Если в доверительный интервал попадает 0 (серая пунктирная линия), то разница не значимая!
- glht (General linear hypotheses and multiple comparisons) для параметрических моделей:

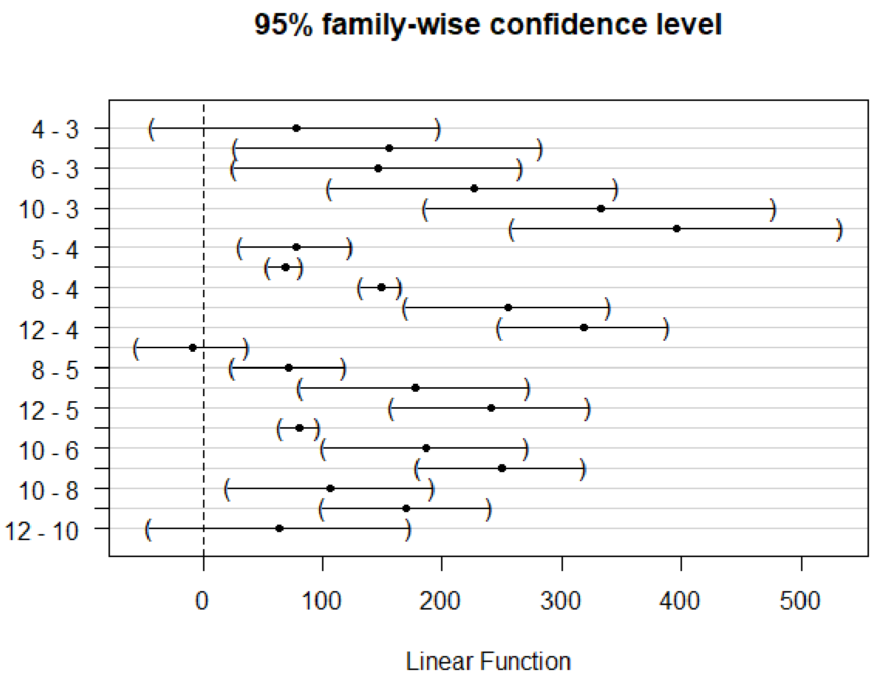
```
glht(model, linfct, alternative =
        c("two.sided", "less",
        "greater"), rhs = 0, ...)
```

Для множественных сравнений:

```
linfct = mcp(…)
```

Пример

```
> aov_model <- aov(Horsepower ~ Cylinders, cars)
> post_test <- glht(aov_model,
+   linfct = mcp(Cylinders = "Tukey")  # или "Dunnett"
+ )
> plot(post_test)
```

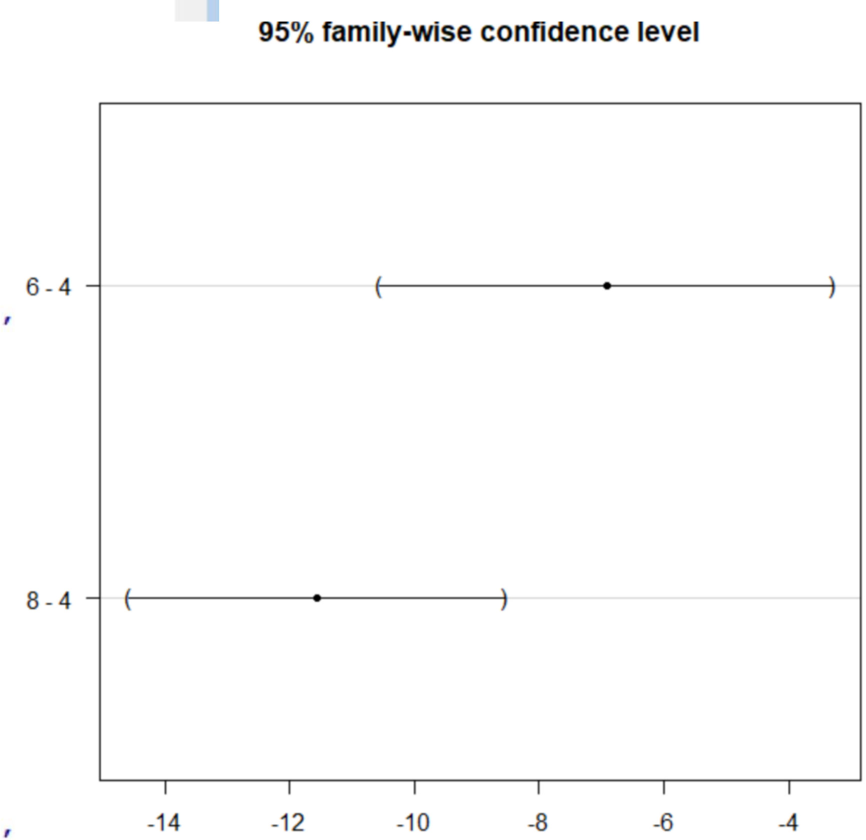



```
> summary(post_test)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Horsepower ~ Cylinders, data = cars)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
4 - 3 == 0     77.515     42.032   1.844  0.44428
5 - 3 == 0    155.000     44.770   3.462  0.00724 **
6 - 3 == 0    145.953     41.989   3.476  0.00679 **
8 - 3 == 0    225.943     42.119   5.364  < 0.001 ***
10 - 3 == 0   332.000     51.291   6.473  < 0.001 ***
12 - 3 == 0   395.667     48.357   8.182  < 0.001 ***
5 - 4 == 0     77.485     16.231   4.774  < 0.001 ***
6 - 4 == 0     68.438      4.704  14.549  < 0.001 ***
8 - 4 == 0    148.428      5.749  25.817  < 0.001 ***
10 - 4 == 0   254.485     29.830   8.531  < 0.001 ***
12 - 4 == 0   318.152     24.444  13.016  < 0.001 ***
6 - 5 == 0     -9.047     16.118  -0.561  0.99670
8 - 5 == 0     70.943     16.453   4.312  < 0.001 ***
10 - 5 == 0   177.000     33.578   5.271  < 0.001 ***
12 - 5 == 0   240.667     28.899   8.328  < 0.001 ***
8 - 6 == 0     79.990      5.421  14.755  < 0.001 ***
10 - 6 == 0   186.047     29.768   6.250  < 0.001 ***
12 - 6 == 0   249.714     24.369  10.247  < 0.001 ***
10 - 8 == 0   106.057     29.951   3.541  0.00544 **
12 - 8 == 0   169.724     24.592   6.902  < 0.001 ***
12 - 10 == 0   63.667     38.230   1.665  0.56838
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Пример

R Console

```
> fcyls<-factor(mtcars$cyl)
> df<-cbind(mtcars,fcyls)
> aov_model<-aov(mpg~fcyls,mtcars)
> summary(aov_model)
            Df Sum Sq Mean Sq F value   Pr(>F)    
fcyls        2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
> post_test<-glht(aov_model,linfct=mcp(fcyls="Dunnett"))
> summary(post_test)

	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = mpg ~ fcyls, data = mtcars)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)    
6 - 4 == 0   -6.921      1.558  -4.441 0.000235 ***
8 - 4 == 0  -11.564      1.299  -8.905 1.71e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
(Adjusted p values reported -- single-step method)
```



Непараметрическая ANOVA - Kruskal-Wallis Test

- Расширение Wilcoxon rank-sum test
  - Обозначим ранги как $r_{ki} = rank(y_{ki})$ и суммы рангов групп как $R_k = \sum_{j=1}^{Nk} r_{kj}$, общее среднее рангов по выборке $\mathrm{R} = (\mathrm{N}+1) / 2$
  - Базовая гипотеза как в ANOVA – групповые ранги совпадают друг с другом, близки к среднему рангу по выборке и значит мало квадратичное отклонение рангов $\sum_{j=1}^{N} n_j (R_j - R)^2$
  - Критерий: $$h = \frac{h^*}{\left(1 - \frac{C}{N(N^2-1)}\right)}$$, где $$h^* = \frac{12}{N(N+1)} \cdot \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i}\right) - 3(N+1)$$
  - C отвечает за корректировку «ничьих», где $\mathrm{C} = \sum_{g=1}^{G} m_g (m_g^2 - 1)$,

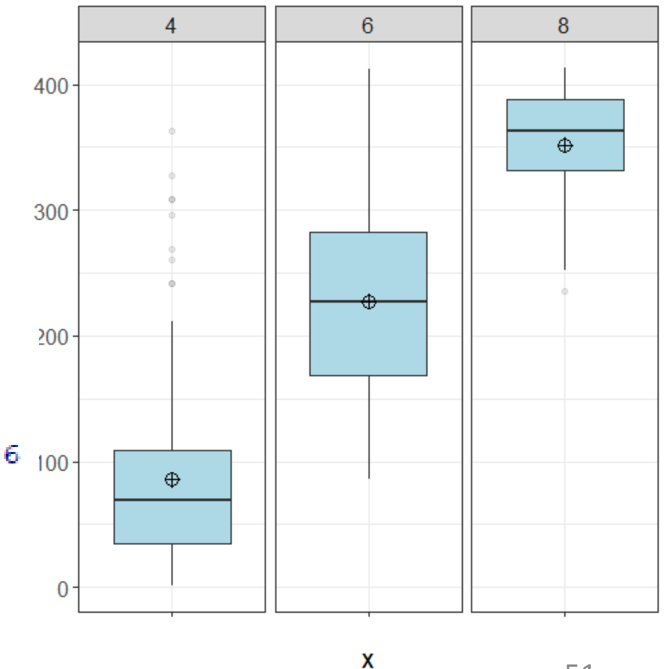
    G – число категорий «ничьих», m_g - число одинаковых рангов в каждой категории
  - Проверка по распределению Хи-квадрат: reject H_0 if $h > \chi^2_{k-1}(\alpha)$

Пример непараметрической ANOVA

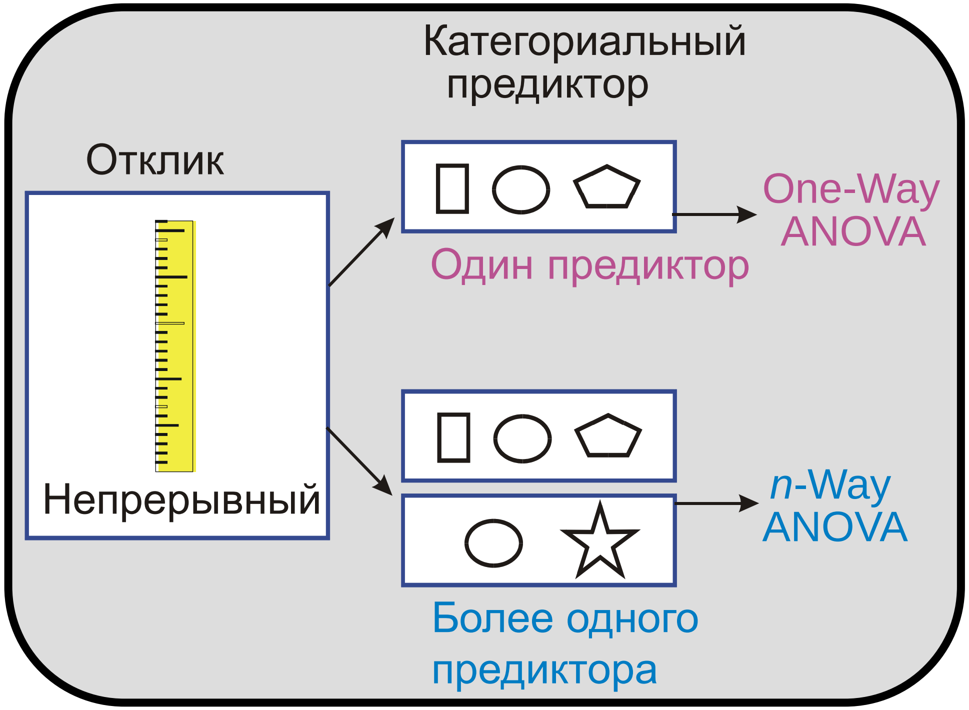
```
> wilcox_test(Horsepower ~ Cylinders, data = sample_3_Cyl, ref.group = "all")
# A tibble: 3 × 9
  .y.        group1 group2    n1    n2 statistic        p    p.adj p.adj.signif
* <chr>      <chr>  <chr>  <int> <int>     <dbl>    <dbl>    <dbl> <chr>
1 Horsepower all    4        413   136    44607  7.02e-25 2.11e-24 ****
2 Horsepower all    6        413   190    35323  4.9 e- 2 4.9 e- 2 *
3 Horsepower all    8        413    87     5354. 7.19e-25 2.11e-24 ****
```

```
> sample_3_Cyl$rank <- rank(sample_3_Cyl$Horsepower)
> ggplot(sample_3_Cyl, aes(x = "", y = rank, group = Cylinders)) +
+ geom_boxplot(fill = "light blue", outlier.alpha = 0.1) +
+ stat_summary(fun=mean, geom="point", shape=10, size=3.5,
+ theme_bw() + theme(text = element_text(size=15)) +
+ facet_wrap(~Cylinders)
```

```
> kruskal.test(Horsepower ~ Cylinders, data = sample_3_Cyl)

        Kruskal-Wallis rank sum test

data:  Horsepower by Cylinders
Kruskal-Wallis chi-squared = 274.95, df = 2, p-value < 2.2e-16
```



Многомерная ANOVA



BloodP = Base Level + Disease + Drug Dose + DrugDose and Disease + Unaccounted for Variation

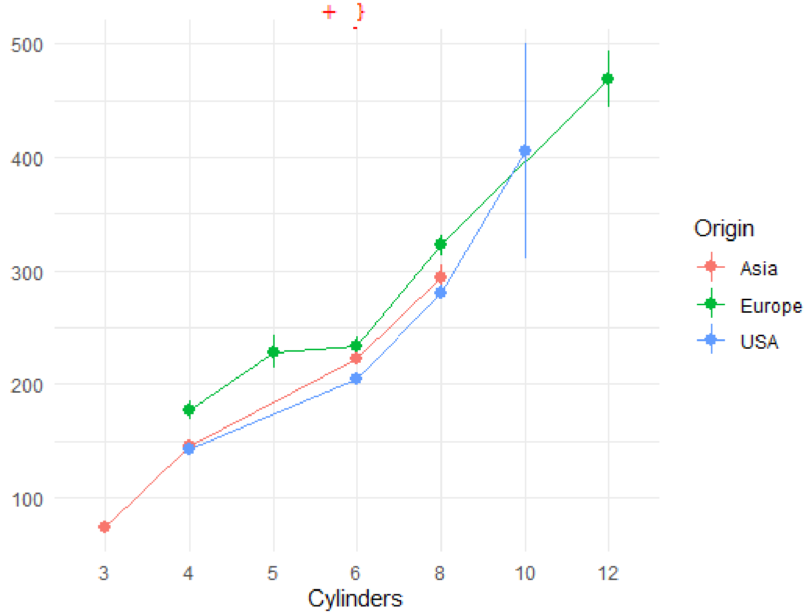
$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Терминология:

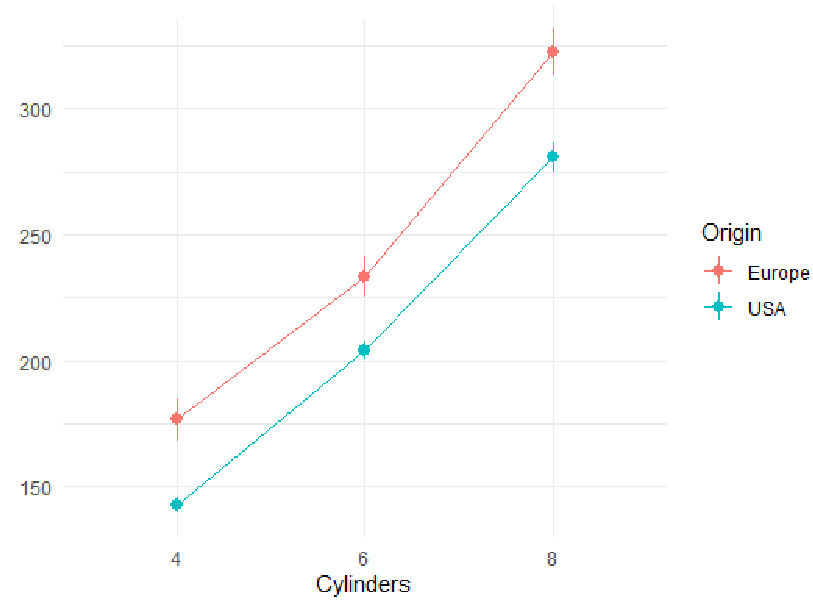
- Модель – математически формализованная связь между предикторами и откликом
- Эффект – ожидаемое изменение в отклике, порождаемое изменением в предикторе
  - Основной эффект – эффект отдельных предикторов (например, x1, x2, x3)
  - Эффект взаимодействия – дополнительный эффект от одновременного изменения двух и более предикторов (например, x1*x2, x1*x2*x3)

Взаимодействующие переменные

```
> plot_interaction<- function(df){
+   na.omit(df) %>% group_by(Cylinders, Origin) %>%
+   summarise(meanloss = mean(Horsepower), se = sem(Horsepower)) %>%
+   ggplot(aes(x = Cylinders, y = meanloss, colour = Origin)) +
+   geom_line(aes(group = Origin)) +
+   geom_pointrange(aes(ymin = meanloss - se, ymax = meanloss + se))
+ }
```




```
> plot_interaction(cars)
```



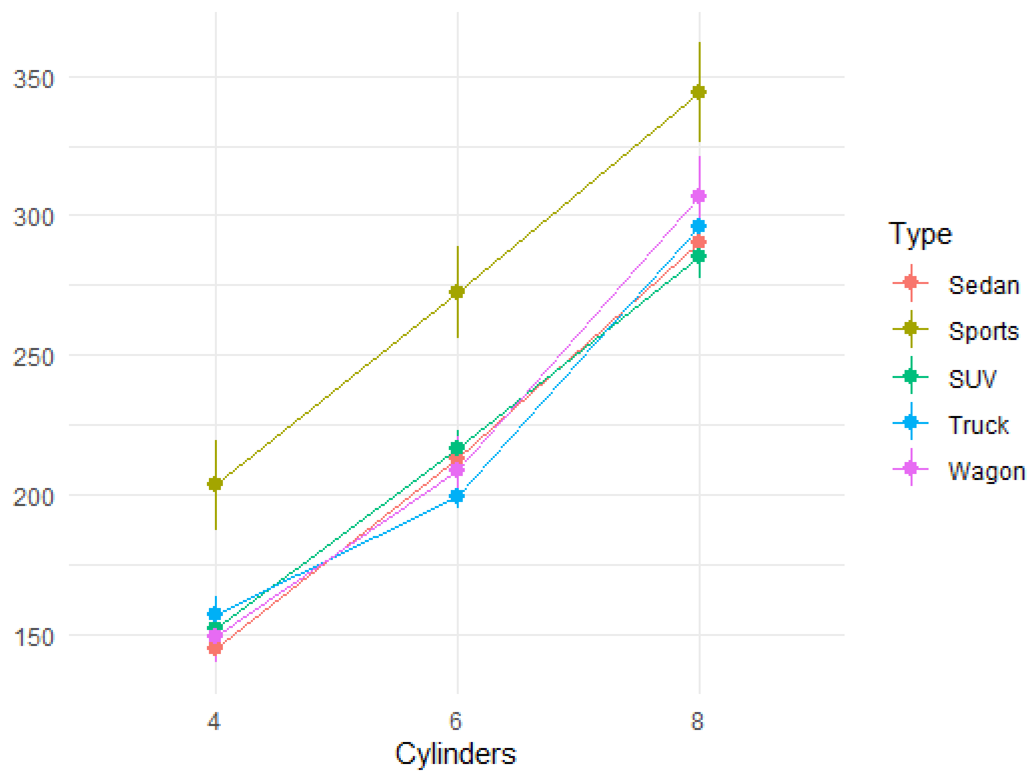

```
> plot_interaction(subset(sample_3_Cyl, Origin != "Asia"))
```

- Строится график среднего отклика со стратификацией по одной из переменных и с группировкой по другой
- Если не пересекаются, то нет взаимодействия и можно упростить модель:

$$\boxed{Y_{ijk} = \mu + \alpha_i + \beta_j + \cancel{(\alpha\beta)_{ij}} + \varepsilon_{ijk}}$$

Задание взаимодействующих переменных и их проверка

```
> summary(aov(Horsepower ~ Cylinders*Type, sample_cyl_type))
                Df  Sum Sq Mean Sq F value Pr(>F)    
Cylinders        2 1160067  580034 406.919 <2e-16 ***
Type             4  129651   32413  22.739 <2e-16 ***
Cylinders:Type   8    5454     682   0.478  0.872    
Residuals      396  564469    1425                   

> plot_interaction(sample_cyl_type)
```



Эвристики для исключения взаимодействующих эффектов (помимо графиков):

- Значение критерия Фишера *F* для члена модели с взаимодействующими эфектами < 2
- Число степеней свободы ошибки < 5 (ErrorDF=Nobs -1 - (ModelDF), где ModelDF = число групп -1)

Пример с взаимодействием переменных

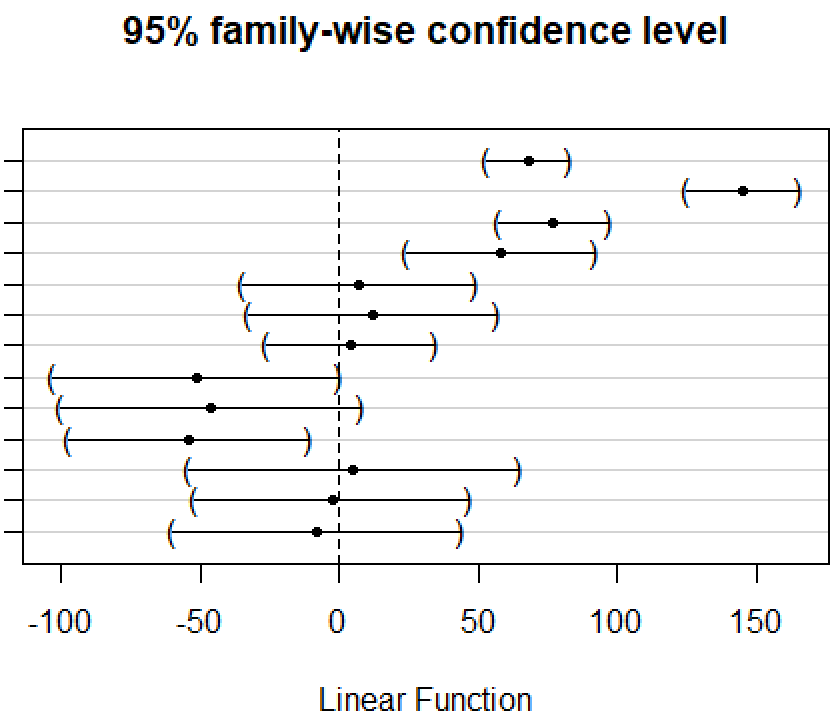
```
> sample_cyl_type <- subset(sample_3_Cyl, Type != "Hybrid")
> sample_cyl_type$id <- rownames(sample_cyl_type)
> summary(lm(Horsepower ~ Cylinders*Type, sample_cyl_type))

Call:
lm(formula = Horsepower ~ Cylinders * Type, data = sample_cyl_type)

Residuals:
     Min       1Q   Median       3Q      Max 
-100.300  -23.240   -0.211   17.767  204.700 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           145.1146     3.8533  37.659  < 2e-16 ***
Cylinders6             68.0021     5.1698  13.154  < 2e-16 ***
Cylinders8            145.0959     7.2360  20.052  < 2e-16 ***
TypeSports             58.2491    12.0180   4.847 1.8e-06 ***
TypeSUV                 6.8854    14.7811   0.466    0.642    
TypeTruck              12.0521    15.8877   0.759    0.449    
TypeWagon               4.3140    10.8011   0.399    0.690    
Cylinders6:TypeSports   0.9343    15.0858   0.062    0.951    
Cylinders8:TypeSports  -4.6739    16.8452  -0.277    0.782    
Cylinders6:TypeSUV     -3.4688    16.6695  -0.208    0.835    
Cylinders8:TypeSUV    -11.6869    17.9104  -0.653    0.514    
Cylinders6:TypeTruck  -26.0576    20.5591  -1.267    0.206    
Cylinders8:TypeTruck   -6.5959    21.1734  -0.312    0.756    
Cylinders6:TypeWagon   -8.9761    16.0663  -0.559    0.577    
Cylinders8:TypeWagon   12.2255    22.5950   0.541    0.589    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.75 on 396 degrees of freedom
Multiple R-squared:  0.6965,	Adjusted R-squared:  0.6857 
F-statistic:  64.9 on 14 and 396 DF,  p-value: < 2.2e-16
```

Пример с взаимодействием переменных

```
> sample_cyl_type$Type <- as.factor(sample_cyl_type$Type)
> post_test <- glht(aov(Horsepower ~ Cylinders*Type, sample_cyl_type),
+   linfct = mcp(Cylinders = "Tukey", Type = "Tukey")  # или "Dunnett"
+ )
```

```
> summary(post_test)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Horsepower ~ Cylinders * Type, data = sampl

Linear Hypotheses:
                            Estimate Std. Error t value Pr(>|t|)
Cylinders: 6 - 4 == 0         68.002      5.170  13.154  < 0.001
Cylinders: 8 - 4 == 0        145.096      7.236  20.052  < 0.001
Cylinders: 8 - 6 == 0         77.094      7.028  10.970  < 0.001
Type: Sports - Sedan == 0     58.249     12.018   4.847  < 0.001
Type: SUV - Sedan == 0         6.885     14.781   0.466  0.99865
Type: Truck - Sedan == 0      12.052     15.888   0.759  0.98125
Type: Wagon - Sedan == 0       4.314     10.801   0.399  0.99944
Type: SUV - Sports == 0      -51.364     18.254  -2.814  0.05040
Type: Truck - Sports == 0    -46.197     19.161  -2.411  0.14177
Type: Wagon - Sports == 0    -53.935     15.212  -3.546  0.00503
Type: Truck - SUV == 0         5.167     21.005   0.246  0.99997
Type: Wagon - SUV == 0        -2.571     17.477  -0.147  1.00000
Type: Wagon - Truck == 0      -7.738     18.422  -0.420  0.99925
```

```
> plot(post_test)
```



Пример без взаимодействующих переменных

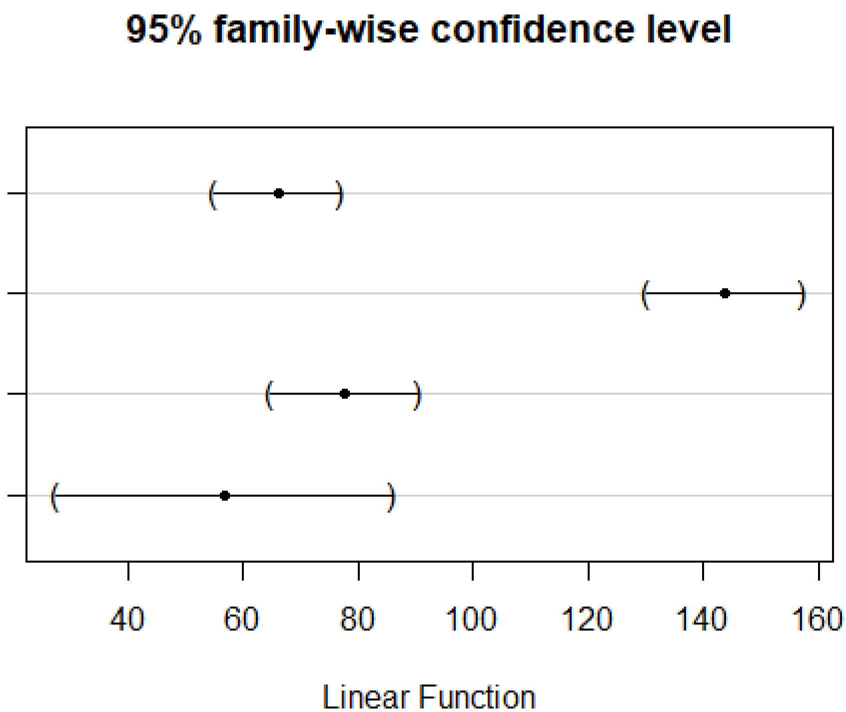
```
> sample_without_interact <- subset(sample_3_Cyl, Type != "Hybrid")
> sample_without_interact$Type = ifelse(sample_without_interact$Type == "Sports",
                                         "Sports", "Other")
```

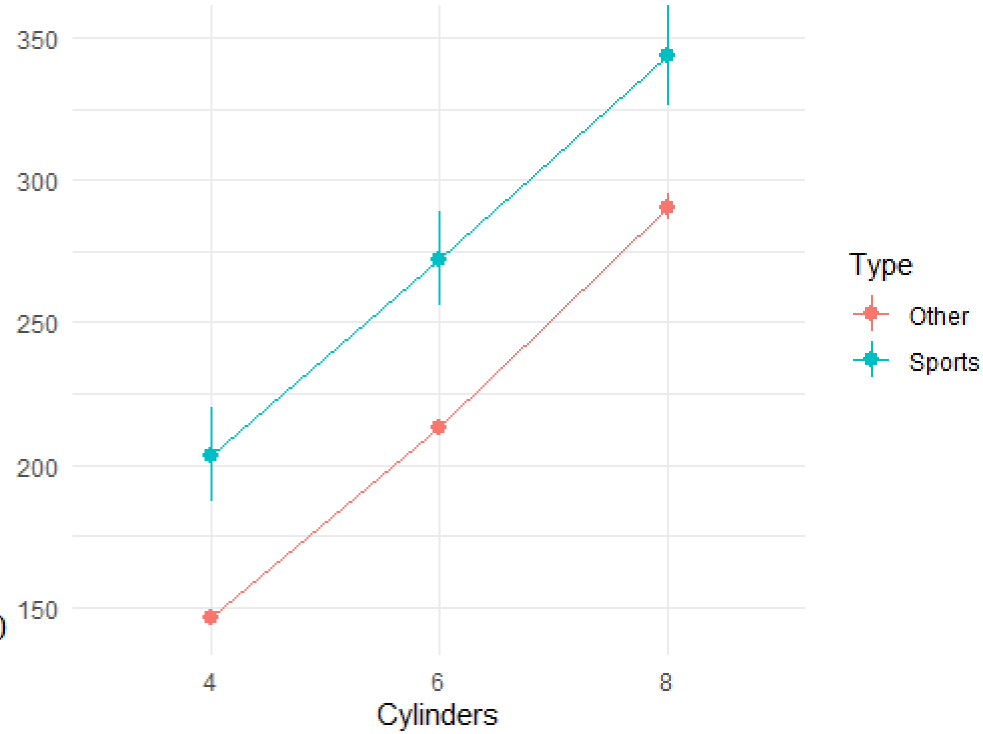
```
> summary(aov(Horsepower ~ Cylinders*Type, sample_without_interact))
                Df  Sum Sq Mean Sq F value Pr(>F)    
Cylinders        2 1160067  580034 412.233 <2e-16 ***
Type             1  129446  129446  91.998 <2e-16 ***
Cylinders:Type   2     272     136   0.097  0.908    
Residuals      405  569857    1407                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lm(Horsepower ~ Cylinders*Type, sample_without_interact))

Call:
lm(formula = Horsepower ~ Cylinders * Type, data = sample_without_interact)

Residuals:
     Min       1Q   Median       3Q      Max 
-100.300  -22.676   -0.676   18.415  204.700 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            146.585      3.382  43.340  < 2e-16 ***
Cylinders6              66.091      4.440  14.884  < 2e-16 ***
Cylinders8             143.757      5.542  25.939  < 2e-16 ***
TypeSports              56.778     11.805   4.810 2.13e-06 ***
Cylinders6:TypeSports    2.845     14.764   0.193    0.847    
Cylinders8:TypeSports   -3.335     16.098  -0.207    0.836    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.51 on 405 degrees of freedom
Multiple R-squared:  0.6936,	Adjusted R-squared:  0.6898 
F-statistic: 183.3 on 5 and 405 DF,  p-value: < 2.2e-16
```

Пример без взаимодействующих переменных

```
> sample_without_interact$Type <- as.factor(sample_without_interact$Type)
> post_test <- glht(aov(Horsepower ~ Cylinders*Type, sample_without_interact),
+   linfct = mcp(Cylinders = "Tukey", Type = "Tukey")  # или "Dunnett"
+ )
```

```
Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
Cylinders: 6 - 4 == 0       66.091      4.440   14.88   <1e-05 ***
Cylinders: 8 - 4 == 0      143.757      5.542   25.94   <1e-05 ***
Cylinders: 8 - 6 == 0       77.666      5.249   14.80   <1e-05 ***
Type: Sports - Other == 0   56.778     11.805    4.81   <1e-05 ***
```





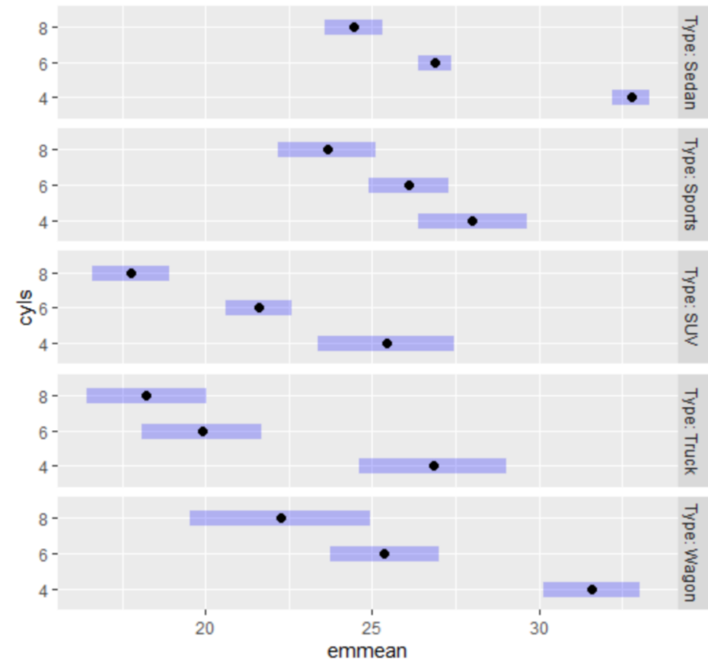
Анализ групповых средних

**emmeans(object, specs, by = NULL, fac.reduce = function(coefs)
apply(coefs, 2, mean), contr, options = get_emm_option("emmeans"),
weights, offset, ..., tran)**

```
> aov_model<-aov(MPG_Highway~cyls*Type,subset(df,cyls %in% c(4,6,8)& Type != "Hybrid" ))
> emmeans(aov_model,~ cyls | Type)
```

```
Type = Sedan:
 cyls emmean    SE  df lower.CL upper.CL
 4      32.8 0.281 396     32.2     33.3
 6      26.9 0.251 396     26.4     27.4
 8      24.4 0.446 396     23.6     25.3

Type = Sports:
 cyls emmean    SE  df lower.CL upper.CL
 4      28.0 0.830 396     26.4     29.6
 6      26.1 0.615 396     24.9     27.3
 8      23.6 0.735 396     22.2     25.1

Type = SUV:
 cyls emmean    SE  df lower.CL upper.CL
 4      25.4 1.040 396     23.4     27.5
 6      21.6 0.502 396     20.6     22.6
 8      17.8 0.587 396     16.6     18.9

Type = Truck:
 cyls emmean    SE  df lower.CL upper.CL
 4      26.8 1.123 396     24.6     29.0
 6      19.9 0.917 396     18.1     21.7
 8      18.2 0.917 396     16.4     20.0

Type = Wagon:
 cyls emmean    SE  df lower.CL upper.CL
 4      31.6 0.735 396     30.1     33.0
 6      25.4 0.830 396     23.7     27.0
 8      22.2 1.376 396     19.5     25.0

Confidence level used: 0.95
```

```
> plot(emmeans(aov_model,~ cyls | Type))
```



CONTRASTS для сравнения «кастомизированных» гипотез

- Для тестов CONTRASTS рассчитывает уровень значимости различия
- Общая идея
  - Записать проверяемую гипотезу в терминах групповых средних
  - Переформулировать проверяемую гипотезу в терминах коэффициентов упрощенной ANOVA модели и подать на вход оператору CONTRASTS
  - Также можно делать через ghlt

Type	Cylinders		
	8	4	
Sports	μ_{11}	μ_{12}	$\mu_{1\cdot}$
Wagon	μ_{21}	μ_{22}	$\mu_{2\cdot}$
Truck	μ_{31}	μ_{32}	$\mu_{3\cdot}$
SUV	μ_{41}	μ_{42}	$\mu_{4\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot\cdot}$

Проверить гипотезу, что у производителей 8 цилиндровых двигателей средняя мощность в группе Sports и Wagon совпадает со средней мощностью в группе Truck и SUV

$$\frac{1}{2}(\mu_{11} + \mu_{21}) = \frac{1}{2}(\mu_{31} + \mu_{41})$$

Проверка кастомизированных гипотез в терминах групповых средних (пример)

- Поскольку

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \text{ и } \frac{1}{2}(\mu_{11} + \mu_{21}) - \frac{1}{2}(\mu_{31} + \mu_{41}) = 0,$$

- то

$$\frac{1}{2}(\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}) - \frac{1}{2}(\mu + \alpha_3 + \beta_1 + (\alpha\beta)_{31} + \mu + \alpha_4 + \beta_1 + (\alpha\beta)_{41}) = 0$$

- И посчитать новые коэффициенты для всех групповых средних:

$$0.5\alpha_1 + 0.5\alpha_2 - 0.5\alpha_3 - 0.5\alpha_4 + 0\beta_1 + 0\beta_2 + 0.5(\alpha\beta)_{11} + 0(\alpha\beta)_{12} + 0.5(\alpha\beta)_{21} + 0(\alpha\beta)_{22} - 0.5(\alpha\beta)_{31} + 0(\alpha\beta)_{32} - 0.5(\alpha\beta)_{41} + 0(\alpha\beta)_{42} = 0$$

- Что приведет к параметрам оператора CONTRAST:
  - Type 0.5 0.5 -0.5 -0.5
    Cylinders *Type 0.5 0 0.5 0 -0.5 0 -0.5 0

Проверка кастомизированных гипотез в терминах групповых средних (пример)

Type	Cylinders		
	8	4	
Sedan	0.5	0	0.5
Wagon	0.5	0	0.5
Truck	-0.5	0	-0.5
SUV	-0.5	0	-0.5

```
> aov_model$coef
      (Intercept)        TypeSports            TypeSUV
      145.1145833        58.2490530          6.8854167
        TypeTruck        TypeWagon               cyls6
       12.0520833         4.3139881         68.0020833
            cyls8 TypeSports:cyls6     TypeSUV:cyls6
      145.0959430         0.9342803         -3.4687500
 TypeTruck:cyls6  TypeWagon:cyls6 TypeSports:cyls8
      -26.0576389        -8.9761093         -4.6738651
    TypeSUV:cyls8  TypeTruck:cyls8   TypeWagon:cyls8
      -11.6868521        -6.5959430         12.2254856
```

```
> tt<-glht(aov_model,linfct=matrix(c(0,1,-1,-1,1,0,0,0,0,0,0,1,-1,-1,1),1))
> summary(tt)

         Simultaneous Tests for General Linear Hypotheses

Fit: aov(formula = Horsepower ~ Type * cyls, data = subset(df, Cylinders %in%
    c("4", "6", "8") & Type %in% c("Sports", "Sedan", "SUV",
    "Wagon", "Truck")))

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0    69.46      26.10   2.661  0.00811 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Выводы по ANOVA

- Нулевая гипотеза = «все средние равны»
- Альтернативная гипотеза: «хотя бы одно среднее отличается»
- Последовательность действий:

1. Постройте описательные статистики и графики
2. Проверьте предположения:
   - Независимость
   - Нормальность ошибки
   - Равенство групповых дисперсий
3. Проверьте *p*-value в табл ANOVA: если меньше заданного уровня значимости alpha, отклоните нулевую гипотезу.
4. Если многомерная, то исследуйте необходимость использования взаимодействующих предикторов